

Journal of Multimedia

ISSN 1796-2048

Volume 7, Number 1, February 2012

Contents

SPECIAL SECTION ON CONNECTED MULTIMEDIA

- Guest Editors' Introduction: Special Section on Connected Multimedia 1
Zhongfei (Mark) Zhang, Zhengyou Zhang, Ramesh Jain, Yueting Zhuang
- Association of Moving Objects Across Visual Sensor Networks 2
Muhammad J. Mirza and Nadeem Anjum
- Faceted Subtopic Retrieval: Exploiting the Topic Hierarchy via a Multi-modal Framework 9
Jitao Sang and Changsheng Xu
- Tour Route Recommendation Begins with Multimodal Classification 21
Xiujun Chen and Qing Wang
-

SPECIAL SECTION ON ISIP 2011

- Guest Editors' Introduction: Special Section on ISIP 2011 31
Fei Yu, Yiqin Lu, Chin-Chen Chang, and Yan Gao
- Speech Separation in the Vehicle Environment Based on FastICA Algorithm 33
Jindong Zhang, Guihe Qin, and Ye Liu
- 97Semantic Analysis of Traffic Video Using Image Understanding 41
Jian Wu, Zhi-ming Cui, Heng-jun Yue, and Guang-ming Zhang
- Research on Video Quality Assessment 49
Chunting Yang, Yang Liu, and Jing Yu
- Tele-Immersive Interaction with Intelligent Virtual Agents Based on Real-Time 3D Modeling 57
Shujun Zhang and Wan Ching Ho
- The Research of Image Encryption Algorithm Based on Chaos Cellular Automata 66
Shuiping Zhang and Huijune Luo
- Improved MFCC Feature Extraction Combining Symmetric ICA Algorithm for Robust Speech Recognition 74
Huan Zhao, Kai Zhao, He Liu, and Fei Yu
- A Watermarking Technique based on the Frequency Domain 82
Huang-Chi Chen, Yu-Wen Chang, and Rey-Chue Hwang
- Image Copy-Move Forgery Detecting Based on Local Invariant Feature 90
Li Jing and Chao Shao
- OWL-S based Service Composition of Threedimensional Geometry Modeling 98
Jiangning Yu, Hongming Cai, Fenglin Bu, and Ailing Liu
-

Guest Editors' Introduction: Special Section on Connected Multimedia

Zhongfei (Mark) Zhang

Computer Science Department, SUNY Binghamton, NY, USA

Zhengyou Zhang

Microsoft Research, Redmond, WA, USA

Ramesh Jain

Department of Computer Science, University of California, Irvine, CA, USA

Yueting Zhuang

College of Computer Science, Zhejiang University, Hangzhou, China

Social media has received extensive attention recently and has become a very popular research area due to its wide spectrum of applications. We note that even though the whole area of social media is very popular in the literature, there are a group of research issues that are related to the social-cultural constraints in the social media study that have not yet received sufficient attention. In this context, we group all these issues together under the umbrella of a new sub-area of social media that we call *connected multimedia*.

Consequently, by connected multimedia, we mean the study of the social and technical interactions among users, multimedia data, and devices across cultures and the explicit exploitation of cultural differences. Hence, connected multimedia involves the three elements – the users, the multimedia data, and the devices – with two perspectives – the social focus and the cultural focus. In short, connected multimedia is about multimedia content and connection across community and cultural boundaries. In comparison with those existing research areas including social media as its super-area and human centered computing, we here emphasize that connected multimedia pays more attention to the cultural difference. The definition of the social side is broader than just national cultures; it possibly includes cultures of groups, disciplines, organizations, communities, ethnicities, religions, and nations. This emphasis distinguishes connected multimedia from all other existing areas such as social media and cross-media, which may either claim to include some of these aspects, among many others, or have different foci.

This special section is organized following the successful first two editions of the workshop on the newly emerged theme of connected multimedia held in Hangzhou, China, in October of 2009 and in Florence, Italy, in October, 2010. After two rounds of rigorous reviews for all the submissions in response to the CFP of this special section, we have finally selected three papers to be included in this special section on connected multimedia.

The paper “Association of Moving Objects across Visual Sensor Networks” by M.J. Mirza and N. Anjum presents an inter-camera trajectory association algorithm for partially overlapping visual sensor networks; the paper addresses issues related to trajectory extraction, representation, and association. The paper “Faceted Subtopic Retrieval: Exploiting the Topic Hierarchy via a Multi-Modal Framework” by J. Sang and C. Xu presents a new framework for multi-modal analysis and retrieval called faceted subtopic retrieval that is tailored to complex queries to the social media data on the Web concerning political and social events or issues. A LDA-like model is developed to exploit the intrinsic topic hierarchy inside the retrieved data. The paper “Tour Route Recommendation Begins with Multimodal Classification” by X. Chen and Q. Wang addresses the problem of location estimation for tourist photos; the paper proposes a solution that begins with classification and exploits explicitly the textual, temporal, geographic, and visual information together for a tour route recommendation. Overall, all the three papers address the theory and application case studies on the general topic of connected multimedia.

While this editorial is co-authored by the guest-editors who also organized the two editions of the workshop on the same topic that actually led to this special section, many people have contributed the ideas that have finally led to the development of this topic of connected multimedia in the literature. Specifically, Noshir Contractor, Alan Hanjalic, Alexander G. Hauptmann, Xian-Sheng Hua, Alejandro (Alex) Jaimes, Ivan Ivanov, Michael S. Lew, Wanqing Li, Ching-Yung Ling, Alexander C. Loui, Jiebo Luo, Michael W. Macy, Nicu Sebe, Qi Tian, Yonghong Tian, Vincent S. Tseng, Qing Wang, Changsheng Xu, Huimin Yu, and Shiwen Yu deserve the credit for contributing their ideas to the development of the literature on this topic. Finally, we would like to thank Dr. Jie Yang, US NSF Program Manager, for the support to the development of this effort and Dr. Jiebo Luo, the EIC of Journal of Multimedia, for the support to publishing this special section. We also acknowledge US NSF (through grant IIS-0956924), Zhejiang University College of Computer Science and Technology, Microsoft Research, and ACM for sponsoring the two editions of the workshop on connected multimedia that led to this special section. This effort of editing this special section is also supported in part by US NSF (IIS-0812114, CCF-1017828) and National Basic Research Program of China (2012CB316400).

Association of moving objects across visual sensor networks

Muhammad J. Mirza and Nadeem Anjum
 Riphah International University
 Hajj Complex Road, I-14 Islamabad (Pakistan)
 {muhammad.javed, nadeem.anjum}@riphah.edu.pk

Abstract— We present a novel inter-camera trajectory association algorithm for partially overlapping visual sensor networks. The approach consists of three steps, namely Extraction, Representation and Association. Firstly, we extract trajectory segments in each camera view independently. These local trajectory segments are then projected on a common-plane. Next, we learn dynamic motion models of the projected trajectory segments using Modified Consistent Akaike's Information Criterion (MCAIC). These models help in removing noisy observations from a segment and hence perform smoothing efficiently. Then, each smoothed trajectory is represented by its curvature. Finally, we use normalized cross correlation, as a proximity measure, to establish correspondence among trajectories that are observed in multiple views. We evaluated the performance of the proposed approach on a simulated and real scenarios with simultaneous moving objects observed by multiple cameras and compared it with state-of-the-art algorithms. Convincing results are observed in favor of the proposed approach.

I. INTRODUCTION

Nowadays video sensors are widespread on airports, train stations and subways for security and surveillance reasons. These sensors are also being used for live sports coverage. Furthermore, TV channels use them to broadcast tragic incidents such as earthquakes and tsunamis. Therefore, video sensors are now common and people are quite accustomed to them e.g., *The Guardian* published a study in 2005 revealed that more than 4.6 million cameras had been installed only in UK for surveillance purposes. Whilst analog CCTV systems were once the norm, they are expensive and require complicated and to some extent difficult installations and constant maintenance. Fortunately, current progress in digital technology has made visual systems far more cost-effective, flexible, and simple to operate. These days, cameras are easy to install and maintain, and hence a network of cameras is now readily available.

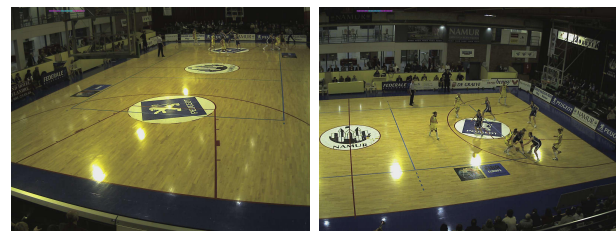


Figure 1. An example of difficulties related to object association in two real, synchronized views. Following difficulties can be observed in the figure: (a) the ball is completely occluded by the players in Left-image; (b) appearance of a player observed in one view is completely different in the other view; (c) within one view, players belonging to same team have similar appearance and motion behaviors and (d) there are considerable shadows for some players in Right-image.

With the increase in number of camera networks, it is becoming more difficult for human operators to analyze the multiple video streams simultaneously; therefore, the problem of effective summarization of the videos is becoming a prime challenge. The association of the objects provides the core for multi-camera information fusion; this means that an object observed in various cameras should be given a unique label throughout the network. There are several factors such as occlusions, shadows and affine projections that make the object association a difficult task [1]. Furthermore, the problem of association becomes more difficult in crowded scenarios, where objects move very close to each other and at times have similar appearance and motion behaviors [2]. Figure 1 shows some of the difficulties that can affect a reliable object association across multiple views.

There are existing approaches that perform object association before tracking [3], [4]. However, we advocate for performing associations after tracking. In this case, each camera tracks objects independently, which rather simplifies the process and also reduces the computational cost of tracking. Then, we project

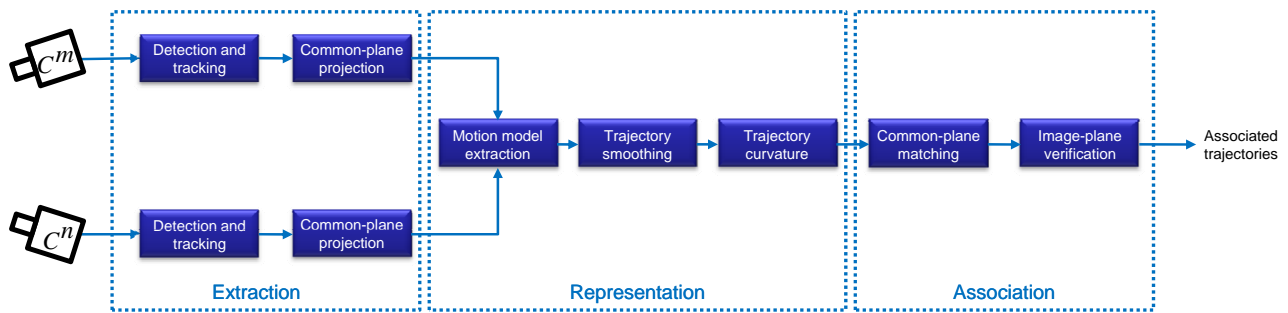


Figure 2. Flow diagram of the proposed approach.

these trajectories on a common-plane using homography transformations; this common-plane can be an image-plane of one of the cameras or a virtual ground-plane [5]. These common-plane trajectories are then utilized to establish object association. To this end, we represent each trajectory by a dynamic motion model, where the goodness of the model is estimated by using Modified Consistent Akaike's Information Criterion (MCAIC) [6]. This provides an in-built mechanism to smooth each trajectory using estimated motion models. Next, curvatures of these smoothed trajectories are considered as a feature vector to represent each trajectory. Finally, we use cross correlation to find the similarity among the trajectories observed in multiple views. The higher the correlation score implies the more the similarity. In case of conflicts, we reproject the trajectories in image-plane and based on Euclidean distance, we select the one corresponding trajectory with minimum distance. A detailed flow diagram of the proposed approach is shown in Fig. 2.

The rest of the paper is organized as follows: Sec. II briefly covers prior works in the field of object correspondence across multiple cameras. Section III provides the detailed description of the proposed approach for object association. Section IV covers the experimental results and finally Sec. V draws conclusions.

II. PRIOR WORK

In this section, the first set of techniques are *supervised*, as they either depend upon the information contained in the training samples or provided manually by users; and the second set of techniques are *unsupervised*, because they do not need training samples or manual selection of the parameters. The algorithms in supervised category are presented by Kettner *et al.* [7], Huang *et al.* [8], Dick *et al.* [9] and Wang *et al.* [10]. On the other hand, the unsuper-

vised algorithms are presented by Kayumbi *et al.* [5], Sheikh *et al.* [11] and Anjum *et al.* [12]. The details of the supervised and unsupervised algorithms, listed above, are provided in rest of this section.

To track people across multiple cameras, Kettner *et al.* [7] presented a Bayesian solution, which requires prior information about the environment and the way people move across it. For tracking cars across two cameras on a highway Huang *et al.* [8] presented a probabilistic approach, where transition times were modeled as Gaussian distributions. Like Kettner *et al.*, it was assumed that the initial transition probabilities were known. To describe patterns of motion for both intra- and inter-camera correspondence, Dick *et al.* [9] use a stochastic transition matrix, where the correspondence between cameras has to be supplied as training data. Wang *et al.* [10] connect trajectories observed in multiple cameras based on their temporal information. The trajectories are considered to be corresponding, if they overlap in time for a empirically pre-selected interval.

Kayumbi *et al.* [5] and Anjum *et al.* [12] establish correspondence between cameras and virtual common-plane. In both approaches, trajectory association is done on the common-plane using several spatio-temporal features. The maximum likelihood for association is calculated by cross correlation of spatio-temporal feature vectors. Another approach in this category is presented by Sheikh *et al.* [11], in their approach, airborne cameras are used with the assumption of the simultaneous visibility of at least one object by two cameras. Taking as input time-stamped trajectories from each view, the algorithm estimates the inter-camera transformations. The maximum likelihood is estimated as a function of the reprojection error. A pair of trajectories is considered as generated from the same object if the reprojection error is minimum.

III. THE PROPOSED APPROACH

A. Trajectory extraction

Let $C = \{C^1, C^2, \dots, C^N\}$ be a set of N partially overlapping synchronized cameras. Let O_n^i represent the i^{th} object observed in C^n . We perform video object extraction (foreground segmentation) using a statistical color change detector and then we associate them across consecutive frames using graph-matching [13]. Let $T_n^i(x, y, t)$ ¹ be the resulting observation (track-point) of O_n^i on location (x, y) at instant t in camera C^n . The trajectory of O_n^i is a set of all observations i.e., $\mathbf{T}_n^i = \{T_n^i(x, y, 0), T_n^i(x, y, 1), \dots, T_n^i(x, y, J_n)\}$, where J_n represent the length of the trajectory.

The image-plane to common-plane (G) projection can be estimated by applying the homography matrix $H_{n,G}$ [14] i.e.,

$$\hat{T}_{n,G}^i(\hat{x}, \hat{y}, t) = H_{n,G} T_n^i(x, y, t), \quad (1)$$

where, $\hat{T}_{n,G}^i(\hat{x}, \hat{y}, t)$ is the local common-plane projection of $T_n^i(x, y, t)$ and $H_{n,G}$ is the homography matrix. $H_{n,G}$ is constructed by selecting control points to establish the image- and common-plane correspondence. However, these local projections result in differences in the overlapping region ($\Omega_{m,n}$) on the common plane. This leads to the requirement of a process which can establish a proximity matrix to associate every p^{th} trajectory to all q^{th} trajectories in $\Omega_{m,n}$.

B. Trajectory representation

In the case of partially overlapping cameras, we need to establish the correspondence between transformed trajectory segments ($\hat{\mathbf{T}}_{n,G}^i$) in the overlapping regions on the common-plane. To find the relative pair-wise similarities for association, we first extract (or select) motion model of each trajectory using Akaike's Information Criterion (AIC) [15]. The criterion is based on the entropy maximization principle. It expresses the motion model in the form of probability distribution and regards fitting a model to the data as estimating the true probability distribution from the data and treats the evaluation and estimation of the model together as one entity by trying to find an estimate which will maximize the expected entropy. The entropy is a natural measure of discrimination between the true and estimated probability distribution; a large value of entropy

¹For the simplicity of notations, superscript i from (x, y) is removed.

means that the distribution is a good approximation to the true distribution. The generalized AIC is given by:

$$AIC(p; \alpha) = -2L_p + \alpha(n)p, \quad (2)$$

where L_p denotes the log-likelihood of the model with p parameters. Furthermore, $\alpha(n)$ is the cost of fitting an additional parameter and n is the number of observations. When there are several competing models, the parameters within the models are estimated by the method of maximum likelihood. The values of AIC's are computed and compared to find a model with the minimum value of AIC. This value is called the minimum AIC estimate (MAICE) and is chosen to be the best model.

Now if we apply AIC to the general regression model of the trajectory $\hat{T}_{n,G}^i$

$$\hat{y} = \hat{x}^1 \theta^1 + \dots + \hat{x}^p \theta^p + \eta, \quad (3)$$

with the assumption that the experimental errors η , are statistically independent, identically normally distributed, with zero mean and variance σ^2 , we have:

$$AIC(p; \alpha) = K(n, \hat{\sigma}) + R_p / \hat{\sigma}^2 + \alpha p, \quad (4)$$

where $K(n, \hat{\sigma}) = n \log(2\pi \hat{\sigma}^2)$ is a constant depending on the marginals of the \hat{x} s. Also, $\hat{\sigma}$ is some estimate of σ^2 . Moreover, $R_p = \sum_{i=1}^n (\hat{y} - \hat{x}^T \hat{\theta}^p)$ is the residual sum of squares with respect to the least squares estimate of parameter vector $\hat{\theta}^p$.

Akaike's choice of $\alpha(n)=2$ is a constant function of n and hence it works well for fixed data set. Objections have been raised that minimizing AIC does not produce an asymptotically consistent estimate of model order [16]. Therefore to get asymptotically consistent estimate, we use Modified Consistent AIC (MCAIC) that penalizes the over-parameterization more strictly. MCAIC can be expressed mathematically as:

$$MCAIC(p) = \mathcal{K} \frac{2 \sum_{i=1}^n \rho(R_{i;p}) + p(\log(n) + 1)}{N}, \quad (5)$$

where \mathcal{K} is an arbitrary constant and $\rho(\cdot)$ is an objective function. In the present work \mathcal{K} is the number of points in the initial window. If the maximum of total observations is not very large for the computing models, then it is better to use the penalty term of AIC instead of MCAIC [6].

In current work, instead of learning single motion model for the entire trajectory, we learn more dynamic models by splitting the trajectory into segments. We learn motion model using MCAIC

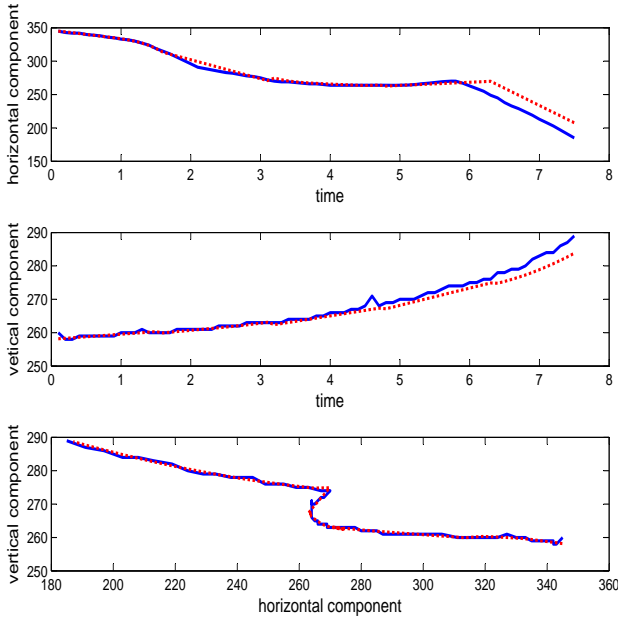


Figure 3. An example of trajectory smoothing using MCAIC: (top,middle) independent smoothing of horizontal and vertical components of the trajectory; and (bottom) complete smoothed trajectory (*dot-line*) superimposed on original trajectory (*solid-line*).

for each trajectory segments. Once the model is learnt, we smooth and resample each trajectory to make them equal sized and represent it with $\tilde{T}_{n,G}^i$. An example of the smoothed trajectory from input trajectory is shown in Fig. 3. To preserve variation information contained in a trajectory, we consider trajectory curvature as a significant feature vector for a trajectory representation. We calculate curvature, at each sample, of the trajectory as:

$$\kappa(t) = \frac{|x'y'' - y'x''|}{(x'^2 + y'^2)^{3/2}}, \quad (6)$$

where $k = 1, \dots, \mathcal{J}$ with \mathcal{J} is length of each resampled trajectory. Also, x' and x'' (y' and y'') are first and second order derivatives along horizontal (vertical) direction. Examples of trajectory curvatures are shown in Fig. 4.

C. Trajectory association

We use cross correlation as a proximity measure to find the similarity of the trajectories due to its robustness against scale variations. If $\kappa_{n,G}^i$ and $\kappa_{m,G}^j$ represent the curvatures of a pair of common-plane trajectories (see Eq. 6) that are originally observed in C^n and C^m then the association matrix is calculated as:

$$A(\tilde{T}_{n,G}^i, \tilde{T}_{m,G}^j) = \varsigma(\kappa_{n,G}^i, \kappa_{m,G}^j), \quad (7)$$



Figure 4. Examples of trajectory curvatures; where, the amount of curvature defines the radius of the marker.

where, ς is the correlation function. A trajectory $\tilde{T}_{n,G}^i$ will be associated to any trajectory $\tilde{T}_{m,G}^j$ for which it has maximum correlation i.e.,

$$D_\Omega = \arg \max_r (A_\Omega(\tilde{T}_{l,G}^j, \tilde{T}_{m,G}^r)) \forall O_m^r \in C^m. \quad (8)$$

In order to have single trajectory segment belong to a physical object in the overlapping region, we need to resolve any conflict situation. For this, we perform matching in image-plane by reprojecting the matched trajectories from the common-plane i.e.,

$$\mathcal{D} = \arg \min_l (d(\tilde{T}_{n,G}^i, \tilde{T}_{m,G}^l)) \forall l = 1, \dots, \mathcal{L}, \quad (9)$$

where $\tilde{T}_{n,G}^i$ and $\tilde{T}_{m,G}^l$ are the trajectories of interest and possible match. Also, \mathcal{L} is the total number of matched trajectories on the common-plane.

IV. EXPERIMENTAL RESULTS

We evaluate the performance of the proposed approach on a simulated as well as real datasets. The simulated dataset is generated using [17] and is consisting of 1816 frames (RGB 24 bit images at 25 frames/sec and 640x480 pixels) from 4 partially overlapping cameras located at different viewpoints

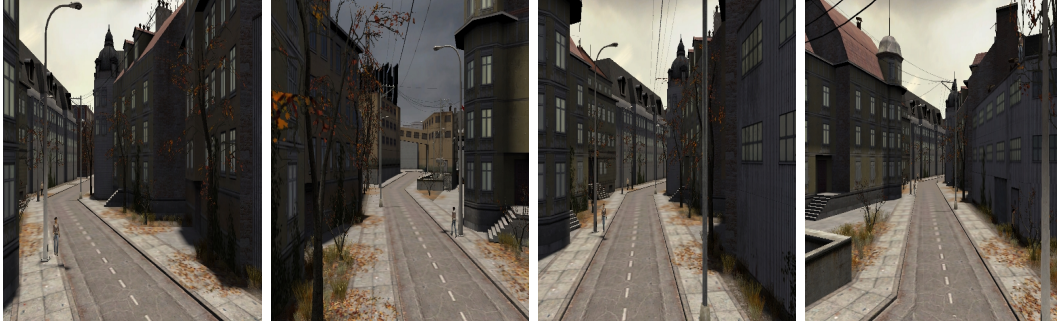


Figure 5. Simulated dataset: Key-frames from each view (left to right: C^1 , C^2 , C^3 and C^4).

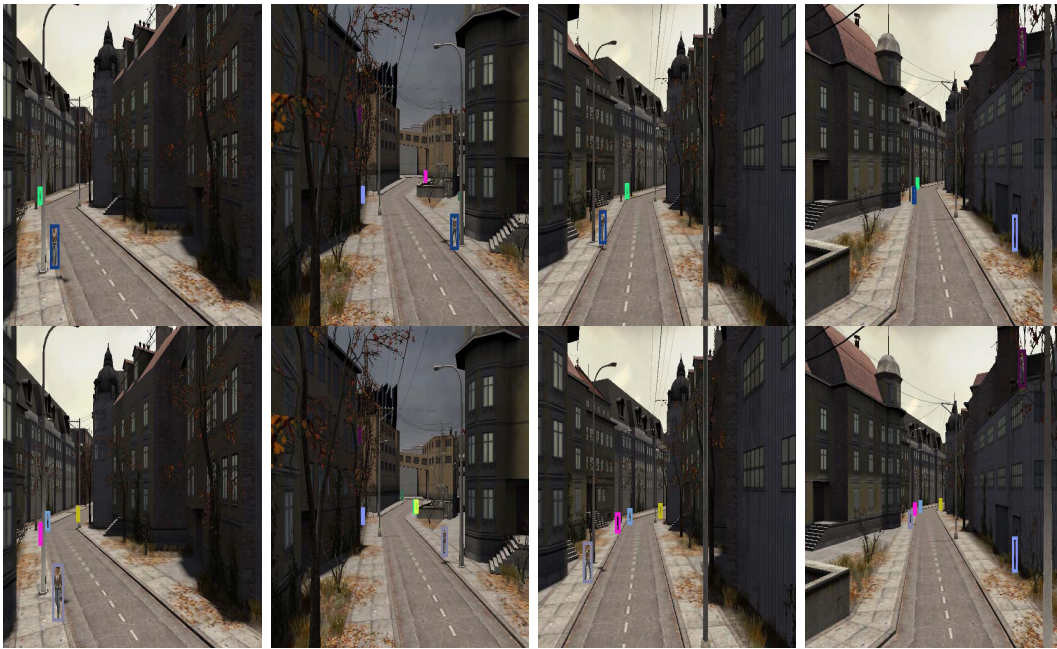


Figure 6. Samples of object association results (color-coded) on *frame-1* and *frame-84* in a partially overlapping network consisting of four simulated cameras. Each column shows the results observed in each camera.

TABLE I.
EVALUATION AND COMPARISON OF OBJECT CORRESPONDENCE RESULTS ON BASKETBALL VIDEO SEQUENCE.

Algorithm	$\Omega_{1,2}$		$\Omega_{1,3}$		$\Omega_{2,3}$		$\Omega_{1,2,3,4}$		Average	
	R	P	R	P	R	P	R	P	R	P
<i>M1</i>	.85	.90	.81	.89	.80	.90	.69	.74	0.79	0.86
<i>M2</i>	.85	.89	.84	.92	.85	.90	.74	.84	0.82	0.89
<i>Proposed</i>	.89	.90	.85	.92	.85	.91	.80	.90	0.85	0.91

(see Fig. 5). The association results of the proposed approach on the simulated network are shown in Fig. 6. It is observed that the proposed approach associated all objects correctly for this video sequence.

We further evaluate the performance of the proposed approach on a real world dataset. The dataset is an indoor basketball video sequence, which consists of 1,000 frames (RGB 24 bit images at 25 frames/sec and 1200x1600 pixels), describing a scene simultaneously recorded by 4 cameras located

at different viewpoints² (see Fig. 7). The closeness of players' movement and similarity in team colors make the association task even more challenging. When acquiring these sequences, no constraints were imposed on objects' trajectories. Figure 8 shows the samples of association results of the proposed approach on basketball sequence. For this dataset, we used visual data to generate the ground truth for association as performed in [12], we perform

²<http://www.apidis.org/Dataset/>

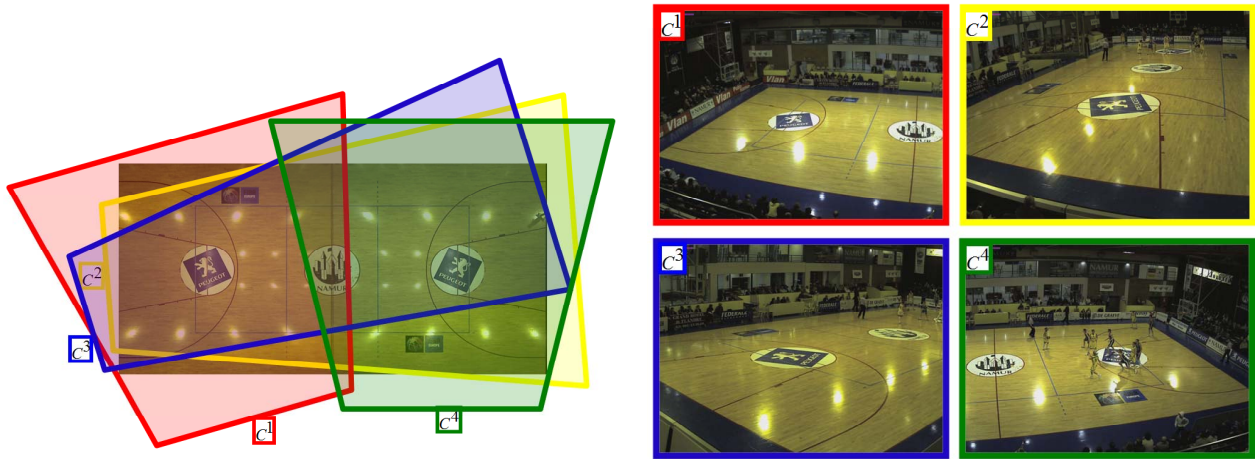


Figure 7. A real basketball video sequence: (Left) configuration of the cameras; (right) key-frames from each view.



Figure 8. Samples of object association results (color-coded) on *frame-1* and *frame-800* in a real basketball video sequence . Each column shows the results observed in each view.

objective evaluation of association and fusion results using *Recall* (R) and *Precision* (P). R is the fraction of accurate associations to the true number of associations. P is the fraction of accurate associations to the total number of achieved associations. Let ξ_{Ω} be the ground truth for pairs of trajectories on the overlapping region Ω and let E_{Ω} be the estimated results. Then R and P are calculated as:

$$R = \frac{|\xi_{\Omega} \cap E_{\Omega}|}{|E_{\Omega}|}, \quad (10)$$

$$P = \frac{|\xi_{\Omega} \cap E_{\Omega}|}{|\xi_{\Omega}|}, \quad (11)$$

where $|\cdot|$ is the cardinality of a set. Table 1 compares the proposed approach with standard dynamic time

warping [18] ($M1$) and the approach presented in [5] ($M2$) in terms of P and R for both sequences. The table compiles the results at camera-pair level as well as at camera-network level. On average, the proposed approach is better by 6% and 5% for R and P respectively, compared to $M1$. Compared to $M2$, the proposed approach outperforms it by 3% and 2% for R and P , respectively. This is because of the more dynamic MCAIC based trajectory representation and built-in verification method. On the other hand, both $M1$ and $M2$ give equal importance to noisy observations as well, which affects the overall association. The results show that on this real world dataset, the proposed approach works accurately for association in both dense and sparse regions.

V. CONCLUSIONS

We addressed an important problem of unsupervised object association after performing tracking in multiple cameras without imposing any physical constraints on placement of the cameras. Initially, local trajectory segments are extracted from each camera, independently. These trajectory segments are then projected on a common-plane using homography transformations. Next to limit the affects of the sensor and/or tracking noise(s) and to smooth the trajectory segments, we used Modified Consistent Akaike's Information Criterion (MCAIC) and each smoothed trajectory segment is then represented with its curvature. Finally, normalized cross correlation is employed as proximity measure to establish correspondence among trajectory segments that are observed in multiple views. In case of conflicts, image-plane verification is employed to resolve conflicting situations. We evaluated the performance of the proposed approach on a simulated and real data and compared the results with two state-of-the-art approaches. It is observed that the proposed approach outperforms both state-of-the-art approached at least by 2% in precision and 3% in recall.

REFERENCES

- [1] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4), 2006.
- [2] F. Daniyal and A. Cavallaro. Content and task-based view selection from multiple video streams. *Multimedia tools and applications*, 46:235–258, 2010.
- [3] M. Taj and A. Cavallaro. Multi-camera track-before-detect. In *Proc. of Intl. Conf. on Distributed Smart Cameras (ICDSC)*, Como, Italy, Aug. 2009.
- [4] A. Sundaresan and R. Chellappa. Multicamera tracking of articulated human motion using shape and motion cues. *Trans. on Image Processing*, 18(9):2114–2126, 2009.
- [5] G. Kayumbi, N. Anjum, and A. Cavallaro. Global trajectory reconstruction from distributed visual sensors. In *Proc. of ACM / IEEE Int. Conference on Distributed Smart Cameras, ICDSC, California, USA*, Sep. 2008.
- [6] M. J. Mirza and K. L. Boyer. An information theoretic robust sequential procedure for surface model order selection in noisy range data. In *Proc. of IEEE Int. Conference on Computer Vision and Pattern Recognition, Champaign, IL, USA*, Jun. 1992.
- [7] V. Kettner and R. Zabih. Bayesian multi-camera surveillance. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), Fort Collins, CO, USA*, June 1999.
- [8] T. Huang and S. Russell. Object identification in a bayesian context. In *Proc. of International Joint Conference on Artificial Intelligence, IJCAI, NAGOYA, Japan*, Aug. 1997.
- [9] A. R. Dick and M. J. Brooks. A stochastic approach to tracking objects across multiple cameras. *Book Series Lecture Notes in Computer Science*, 3339/2005:160–170, Nov. 2004.
- [10] X. Wang, K. Tieu, and W.E.L. Grimson. Correspondence-free multi-camera activity analysis and scene modeling. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CVPR, Alaska, USA*, Jun. 2008.
- [11] Y.A. Sheikh and M. Shah. Trajectory association across multiple airborne cameras. *Trans. on Pattern Analysis and Machine Intelligence*, 30(2):361–367, Feb. 2008.
- [12] N. Anjum and A. Cavallaro. Trajectory association and fusion across partially overlapping cameras. In *Proc. of Intl. Conf. on Advanced Video and Signal based Surveillance, Genoa, Italy*, Sep. 2009.
- [13] M. Taj, E. Maggio, and A. Cavallaro. Multi-feature graph-based object tracking. In *CLEAR, Springer LNCS 4122*, pages 190–199, Southampton, UK, Apr. 2006.
- [14] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Second ed. Cambridge University Press, UK, 2004.
- [15] A. Hirotsugu. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716723, 1974.
- [16] R. L. Kashyap. Optimal choice of ar and ma parts in autoregressive moving average models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4(4):99104, 1982.
- [17] G. Taylor, A. Chosak, and P. Brewer. Ovvv: Using virtual worlds to design and evaluate surveillance systems. In *Proc of IEEE int conf on computer vision and pattern recognition*, 2007.
- [18] R. Turetsky and D. Ellis. Ground-truth transcriptions of real music from force-aligned midi syntheses. In *4th International Symposium on Music Information Retrieval, Baltimore, USA*, Oct. 2003.

Faceted Subtopic Retrieval: Exploiting the Topic Hierarchy via a Multi-modal Framework

Jitao Sang^{1,2}, Changsheng Xu^{1,2}

¹National Lab of Pattern Recognition, Institute of Automation, CAS, China

²China-Singapore Institute of Digital Media, Singapore

Email: {jtsang, csxu}@nlpr.ia.ac.cn

Abstract—The overwhelming amount of web videos posted on the social media websites make effective browsing and search a challenging task. The user-provided metadata, has been proved useful in large-scale video organization and retrieval. Search result clustering, which utilizes the associated metadata to cluster the returned results into semantic groups according to its involved subtopics, has shown its advantages. Most of the existing works on search result clustering are devoted to solving the *ambiguous* problem resulted from general queries. In this paper, we propose the problem of *faceted subtopic retrieval*, which focus on more complex queries concerning political and social events or issues. Hierarchical topic model (hLDA) is adapted to exploit the intrinsic topic hierarchy inside the retrieved collections. Furthermore, this paper offers a new perspective of multi-modal video analysis by exploring the pairwise visual cues deriving from duplicate detection for constrained topic modeling. We modify the standard hierarchical topic model by integrating: 1) query related Supervision knowledge (ShLDA) and 2) duplicate Relation constraints (RShLDA). Carefully designed experiments on web-scale video dataset validate the proposed method.

Index Terms—subtopic retrieval, multi-modal analysis, search result clustering, hierarchical topic model, social media, connected multimedia

I. INTRODUCTION

With the development of multimedia technology and increasing proliferation of social media in Web 2.0, an overwhelming volume of images and videos have been posted to the media sharing websites, making effective browsing and searching a challenging task. It becomes extremely difficult for users to find the information they need. The quality of search engines are often far from satisfactory due to various reasons. Firstly, although ranked lists are still popular ways for organizing the search results, it is highly inefficient since users have to painstakingly browse through the long list to judge whether the results match their requirements. Secondly, a majority of the queries tend to be short, non-specific and imprecise, which makes the relevance-based rank results even unreliable.

One approach to address the above issues is search result clustering, which clusters and visualizes the search results into semantically consistent groups. The main advantage is that it favors systematic exploration and helps users get a quick overview of search results, which

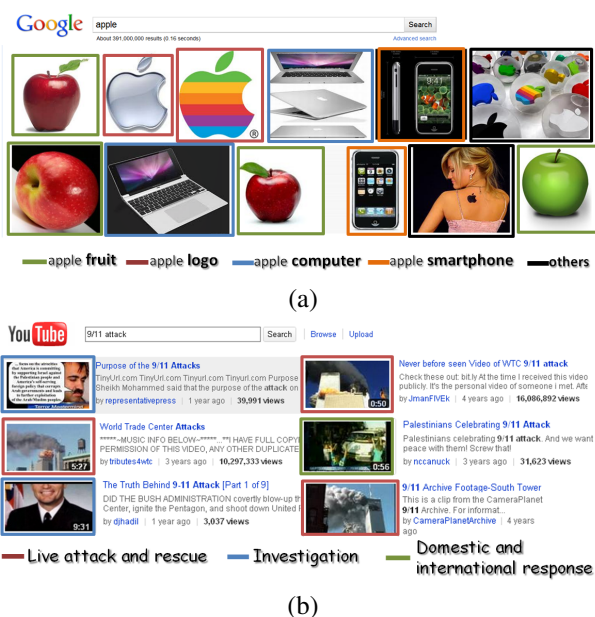


Figure 1. Example queries. (a) Ambiguous query and subtopics in its search results from Google’s image search (b) Faceted query and subtopics from Youtube

decreases the effort required to locate relevant information. Moreover, when the queries are short, it is usually involved with multiple aspects of potential interest and the relevance criterion alone is not sufficient, which aggravate the limitation of the rank principle.

In search result clustering, ideally, one cluster corresponds to one subtopic of the query-related topic. Actually, according to different queries, there are two categories of subtopics. The first is called *ambiguous* subtopic, where the queries have multiple interpretations and each subtopic corresponds to one independent interpretation (see an example in Fig.1(a)). The ambiguous queries most involve general objects or names. Most of the previous search result clustering methods [1]–[5] are devoted to solving this ambiguous problem. Another is called *faceted* subtopic, where the queries have unique interpretation with each subtopic covering different facet of the query (see an example in Fig.1(b)). Most of the queries concerning political and social events or issues belong to this category, which is the focus of this work. In this paper we define the problem of **faceted subtopic retrieval** as

clustering and visualizing the search results into faceted subtopics, and we present a novel multi-modal framework combining the user-provided textual metadata and the video visual content. From the perspective of utilizing the user-provided metadata, this work can be viewed as one implementation of the new research field - connected multimedia, which is devoted to bridging the semantic gap in multimedia analysis by the aid of user involvement and cultural constraints [6].

We observe that there exists an intrinsic hierarchical structure in the search results collection returned from faceted queries. Simply taking a glance at the example in Fig.1(b), we find that almost all the returned videos contain words like ‘9/11’, ‘attack’, ‘terrorism’, ‘WTC’, etc. This is reasonable and implies that although diverse topics are involved in the retrieved video collection, they usually share one common topic referred in query. Each retrieved video can be viewed as a combination of the shared topic (“9/11 terrorist attack”) and one subtopic (which can be enumerated as “Live attack and rescue video”, “Domestic and international response afterwards”, “Investigation” and “The Else: long-term effect and memorial”). Delighted from this, we extend the hierarchical Latent Dirichlet Allocation (hLDA, [7]) to exploit a two-level topic tree in the retrieved video collection and cluster the collected videos into the leaf-level subtopics.

Furthermore, we provide a novel multi-modal framework to fuse textual features (user-provided metadata, e.g., title, description, tags) and visual features (video content). Although many efforts have been targeted at building concept detector from low-level visual features, the performance of high-level detector is far from satisfactory due to the notorious semantic gap. These years near-duplicate cluster detection has attracted extensive research attentions. Several effective methods have been proposed and the detection precision is acceptable. However, typically the near-duplicate detection is based on keyframe extraction and matching from the low-level visual features. The video clips in the same near-duplicate cluster are basically derived from the same original copy and are not sufficient to represent the semantic subtopics. Instead of directly constructing the clusters from the duplicate detection, in this paper we assume that, if two videos are near-duplicate, they should be semantically clustered together and belong to the same subtopic. We propose to employ the video content for near-duplicate detection and integrate the obtained duplicate correlation between every two videos as the pairwise visual constraints on top of textual metadata based topic modeling. To this end, we extend the traditional hierarchical topic model to its relational version.

The main contributions of this paper are summarized as two aspects.

- We propose the problem of faceted subtopic retrieval and adapt the method of hierarchical topic model to exploit the latent hierarchical structure. Furthermore, query is incorporated to guide the topic hierarchy discovery and a supervised extension (ShLDA) to

the standard hierarchical topic model is introduced.

- A novel multi-modal framework is presented. The pairwise duplicate relations are utilized as the visual constraints onto the textual feature based topic modeling. We further extend the ShLDA to its relational version, relational ShLDA (RShLDA).

The rest of the paper is structured as follows. In section II, we review related work on search result clustering and multi-modal video analysis. The overview of our framework is described in section III, whereas the details are presented in section IV and section V. Section VI reports our experimental results on web video dataset and section VII concludes our work.

II. RELATED WORK

In this section, we review the previous work on search result clustering and multi-modal video analysis. Their relations with our work are also discussed.

A. Search Result Clustering

Search result clustering, which clusters the raw returned documents into different semantic groups has been investigated in text retrieval [1], [2], image retrieval [3], [4] and video retrieval [5]. Comparing with the conventional rank list, the labeled clusters allow better topic understanding and favor systematic exploration of the search results.

Most of the previous search result clustering methods are devoted to solving the *ambiguous* problem resulted from non-specific queries. The queries most involve general objects or names, and the cluster labels (ambiguous subtopics) correspond to alternative interpretations of the query. For example, query ‘*apple*’ with interpretation of computer, ipod, logo and fruit [3]; query ‘*sting*’ with interpretation of musician, wrestler and film [5]. In this paper we focus on more complex queries concerning political and social events or issues. These queries have unique interpretation and the subtopics are diverse facets of the query-corresponding events (e.g. query of ‘9/11 attack’) or different viewpoints on controversial issues (e.g. query of ‘*abortion*’ with opposing viewpoints of ‘pro-life’ and ‘pro-choice’). The existing methods are not adaptable to the problem of faceted subtopic retrieval as they take no consideration on the intrinsic hierarchical topic structure.

Note that the term of ‘faceted retrieval’ was also defined in [8]. However, their work is devoted to the TREC diversity task which aims to return a ranked list providing as much coverage for a query as possible [9]. In their case, “faceted” indicates each document may cover multiple facets of a topic.

B. Multi-modal Video Analysis

Web videos carry rich textual metadata as well as video content, which together provide important clues for video analysis and topic mining. The method of combining multiple modality features, however, is not trivial. One straight approach is to concatenate features from each

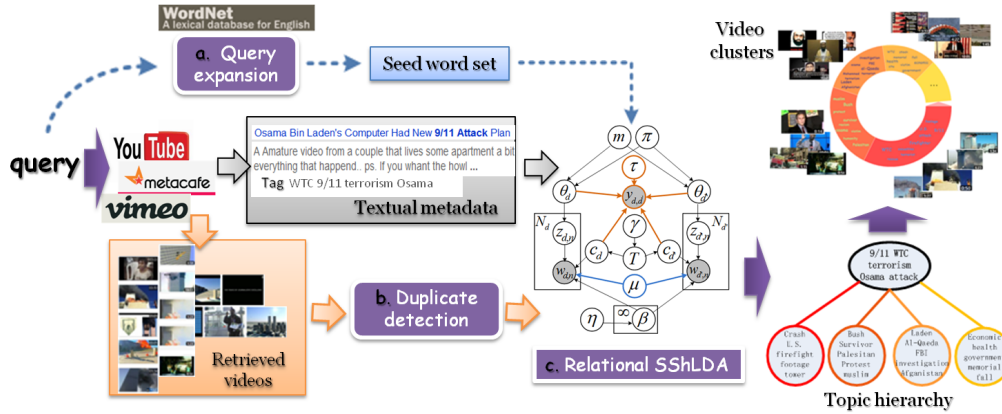


Figure 2. The proposed framework.

modality to make a long feature vector. This method will suffer from the “curse of dimensionality” [10] and concatenating the multidimensional features simply does not scale up. Another approach is to build individual models for each modality, and combine them for final decision [11], [12]. This approach lacks solid theories and the unsatisfactory performance of the visual concept detector make the learnt models unreliable and hardly generalizable. Recently, machine learning algorithms of graph-based [13], co-clustering [14] and multiple kernel learning [15] are proposed to combine multiple types of heterogeneous data. However, the expensive cost of feature processing is not acceptable for the problem of search result clustering.

In this paper, we employ the visual video content by duplicate detection and integrate the duplicate link into the topic modeling process. Effective duplicate detection methods have been proposed and the precision is satisfactory. A statistic found that there are 20%-30% of near duplicate videos in the web collections [16], which indicates that there are abundant near duplicate link available for topic mining in web videos.

III. FRAMEWORK

In this paper, we propose a relational hierarchical topic model based multi-modal framework for faceted subtopic retrieval. This work attempts to organize the search results from web video search engines into semantic clusters, with each cluster corresponding to one subtopic of the faceted query. The framework is shown in Fig.2. The input of our algorithm is the submitted queries, the web videos and associated textual metadata collected from social media websites (e.g., Youtube, Metacafe, Vimeo, etc.). The output is the generated video clusters as well as the topic hierarchy. The framework contains three components, query expansion, duplicate detection and hierarchical topic model based topic hierarchy discovery. For query expansion, we employ association mining as well as WordNet conceptual relation to expand the query words, resulting in a seed word set. The seed word set is viewed as the supervision information (we refer it as query-root-topic knowledge) for the latent root topic and will later be integrated into the topic hierarchy discovery

process. The retrieved videos are sent to the duplicate detection module to find the duplicate clusters, which are then utilized as the visual constraint to guide the inference of the topic hierarchy. We note that the duplicate detection can be performed offline at the time the videos are uploaded. The details of query expansion and duplicate detection are presented in section IV. The extension to the traditional hierarchical topic model, RShLDA, is the core of our framework. RShLDA utilizes the supervision information of query-root-topic and the observed document-level duplicate link to help discover the topic hierarchy. After probabilistic inference, each video is assigned a single path from the root node to a leaf node. The videos assigned to the same path will be grouped together to form a semantic cluster and the subtopics in the leaf node constitute the description for the corresponding video clusters.

IV. TOPIC HIERARCHY DISCOVERY VIA HIERARCHICAL TOPIC MODELLING

We begin this section by briefly reviewing LDA and the standard hLDA. Then we introduce our extensions to hLDA, ShLDA and RShLDA. The inference algorithms are also discussed. In the following, we will describe the models using the original terms “documents” and “words” as used in the topic model literatures (in our case, each video corresponds to one document).

A. LDA and hLDA

Suppose we have a corpus of M documents, $\{w_1, w_2, \dots, w_M\}$ containing words from a vocabulary of V terms. Further we assume that the order of words in a particular document is ignored. This is a ‘bag-of-words’ model.

LDA: Latent Dirichlet Allocation model [17] assumes that documents are generated from a set of K (K needs to be predefined) latent topics. In a document, each word w_i is associated with a hidden variable $z_i \in \{1, \dots, K\}$ indicating the topic from which w_i was generated. The probability of word w_i is expressed as

$$P(w_i) = \sum_{j=1}^K P(w_i|z_i = j)P(z_i = j) \quad (1)$$

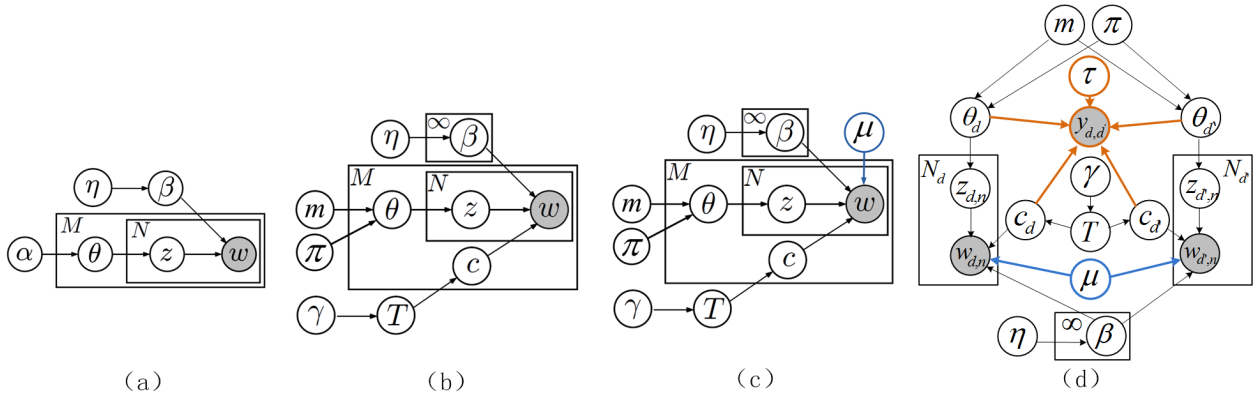


Figure 3. The graphical representation for the topic models: (a) LDA (b) hLDA (c) ShLDA and (d) RShLDA. The parameter μ controls the strength of query-root-topic supervision. The variable $y_{d,d'}$ indicates whether the two videos are near-duplicates. The proposed ShLDA and RShLDA differs from the standard hLDA in the way the words \mathbf{w} are generated.

where $P(w_i|z_i = j) = \beta_{ij}$ is a probability of word w_i in topic j and $P(z_i = j) = \theta_j$ is a document specific mixing weight indicating the proportion of topic j . LDA treats the multinomial parameters β and θ as latent random variables sampled from a Dirichlet prior with hyperparameters α and η respectively. The corresponding graphical model is shown in Fig.3(a).

hLDA: The LDA model described above has a flat topic structure. Each document is a superposition of all K topics with document specific mixture weights. The hierarchical LDA model organizes topics in a tree of fixed depth L . Each node in the tree has an associated topic and each document is assumed to be generated by topics on a single path from the root to a leaf through the tree. Note that all documents share the topic associated with the root node, this feature of hLDA is consistent with the characteristics of search result collection we mentioned in Section I. The merit of the hLDA model is that both the topics and the structure of the tree are learnt from the training data. This is achieved by placing a nested Chinese restaurant process (nCRP) [18] prior on the tree structure. nCRP specifies a distribution on partitions of documents into paths in a fixed depth L -level tree. To generate a tree structure from nCRP, assignments of documents to paths are sampled sequentially, where the first document forms an initial L -level path, i.e. a tree with a single branch. The probability of creating novel branches is controlled by parameter γ , where smaller values of γ result in a tree with fewer branches. In hLDA, each document is assumed drawn from the following process:

- i. Pick a L -level path \mathbf{c}_d from the nCRP prior: $\mathbf{c}_d \sim nCRP(\gamma)$.
- ii. Sample L -dimensional topic proportion vector $\theta_d \sim GEM(m, \pi)$.
- iii. For each word $w_{d,n} \in \mathbf{w}_d$:
 - (a) Choose level $z_{d,n} \sim Discrete(\theta_d)$;
 - (b) Sample a word $w_{d,n} \sim Discrete(\beta_{\mathbf{c}_d|z_{d,n}})$, which is parameterized by the topic in level $z_{d,n}$ on the path \mathbf{c}_d .

The corresponding graphical model is shown in Fig.3(b).

\mathcal{T} denotes the tree structure generated from the nCRP process, \mathbf{c} is the selected path of documents, hyperparameter η is the same with that in LDA and control the smoothing/sparsity of topic-word distribution¹, while GEM parameters $\{m, \pi\}$ reflect the stick-breaking distribution and control the document-topic allocation. Further details of hLDA can be found in [7]. When we utilize hierarchical topic model for the video clustering task, one subtopic corresponds to one cluster. The cluster membership of each video is decided by its posterior path assignment \mathbf{c}_d . The cluster videos are sorted by their proportion on the subtopic as computed by:

$$\frac{\sum_{w_{d,n} \in \mathbf{w}_d} |z_{d,n} = 2|}{N_d} \quad (2)$$

where $|\cdot|$ is indicator function and the numerator accumulates the number of words allocated at the leaf level, N_d denotes the word number.

B. ShLDA

To incorporate the query-root-topic knowledge into the hierarchical topic modeling, we propose the supervised hLDA model (ShLDA). The supervision information we add is the seed word set derived from query expansion, \mathcal{S} . We jointly model the documents and the seed word set, in order to guide the discovery of topic hierarchy so that the words in the seed word set will have high probability in the root topic and low probability in subtopics.

We first explain how query-root-topic knowledge can be incorporated into the topic modeling process. In the standard hLDA, the topic level allocation $z_{d,n}$ for word n in document d is a latent variable and needs to be inferred through the model learning process. Assume we have the supervised information of $z_{d,n}$, i.e. the topic level allocation for a given word in a given document. We denote it as hard constraint when the seed set words are restricted to be shown only in the root topic. In practical applications, each word tends to be generated from every topic with different probabilities. Therefore, we relax the strong assumption that seed words only be generated from

¹ ∞ denotes that there no constraint for the number of topics.

the root topic. Instead of providing topic level allocation $z_{d,n}$ for each seed word, we modify the generative process of standard hLDA so that sampling seed words from root topic and subtopics will have different probabilities.

Specifically, the proposed ShLDA differs from hLDA in the way $w_{d,n}$ is generated. The generative process of ShLDA is:

- i. Pick a L -level path \mathbf{c}_d from the nCRP prior: $\mathbf{c}_d \sim \text{nCRP}(\gamma)$.
- ii. Sample L -dimensional topic proportion vector $\theta_d \sim \text{GEM}(m, \pi)$.
- iii. For each word $w_{d,n} \in \mathbf{w}_d$:
 - (a) Choose level $z_{d,n} \sim \text{Discrete}(\theta_d)$;
 - (b) Sample a word $w_{d,n} \sim \text{Constraint}(\mu, z_{d,n}) \cdot \text{Discrete}(\beta_{\mathbf{c}_d} | z_{d,n})$

The corresponding graphical model is shown in Fig.3(c). $\text{Constraint}(\mu, z_{d,n})$ is the soft constraint function defined as follows:

$$\text{Constraint}(\mu, z_{d,n}) = \begin{cases} \mu \delta(w_{d,n} \in \mathcal{S}) + 1 - \mu, & z_{d,n} = 1, \\ \mu \delta(w_{d,n} \notin \mathcal{S}) + 1 - \mu, & \text{else.} \end{cases} \quad (3)$$

where $\mu (0 \leq \mu \leq 1)$ is the strength parameter of the supervision and $\delta(\cdot)$ is an indicator function with value 1 if the condition is satisfied and 0 otherwise. $\mu = 0$ reduces to standard hLDA and $\mu = 1$ recovers the hard constraint.

The formulation in Eq.3 provides us a flexible way to insert a prior domain knowledge into the inference of latent topics with different definitions of the constraint function, e.g. with prior information on the latent subtopics, \mathcal{S} can be set independently for the specific subtopic.

C. RShLDA

Besides the observed words, the links between documents are also important cues for topic modeling. For each pair of documents d and d' , the results of duplicate detection are viewed as the observed link and modeled as a binary random variable $y_{d,d'}$ conditioned on their contents. Duplicate link variable $y_{d,d'} = 1$ if d and d' are near-duplicate, and $y_{d,d'} = 0$ otherwise. In this sense, we propose RShLDA, a relational extension to the proposed ShLDA. RShLDA is a model of data composed of documents and links between them. It embeds these data in a latent topic space which explains both the words of the documents and how they are connected.

In RShLDA, each document is first generated as in ShLDA. The duplicate links between every two documents are then modeled as binary variables. It assumes that a set of observed documents $w_{1:D,1:N_d}$ and binary duplicate links between them $y_{1:D,1:D}$ are generated by the following process ²:

² For $w_{1:D,1:N_d}$, $1 : D$ and $1 : N_d$ denote the range of its subscripts, where D is the number of documents and N_d is the number of words in document d . Similar representation is used in $y_{1:D,1:D}$.

- i. For each document d : performs the same generative process as ShLDA.
- ii. For each pair of documents d, d' :

- (a) Draw binary duplicate link indicator

$$y_{d,d'} \sim \psi(\cdot | \mathbf{c}_d, \mathbf{c}_{d'}, \theta_d, \theta_{d'}, \tau)$$

The function $\psi(\cdot)$ is the link probability function that defines a distribution over the duplicate link between two documents. It is dependent on the path assignments $\mathbf{c}_d, \mathbf{c}_{d'}$ and the topic proportions $\theta_d, \theta_{d'}$:

$$\psi(y_{d,d'} = 1) = \begin{cases} \sigma(\tau^T(\theta_d \circ \theta_{d'})), & \text{if } \mathbf{c}_d = \mathbf{c}_{d'} \\ 0, & \text{else.} \end{cases} \quad (4)$$

where τ is coefficient parameter, \circ denotes the element-wise product and $\sigma(\cdot)$ is the sigmoid function. This formulation ensures that the same latent topic assignments used to generate the content of the documents also generate the link structure. We can see that only when two documents are assigned to the same path (subtopic), the binary variable has the probability to be one. The more similar the topic proportions of the two documents are, the more likely the duplicate link exists. This is consistent with our assumption in section I. Fig.3(d) illustrates the graphical model for RShLDA for a single pair of documents. Note that the full model contains D^2 duplicate link variables and is difficult to illustrate.

D. Inference

The central computational problem in Bayesian topic modeling is the posterior inference, which inverts the generative process to estimate the conditional distribution of the latent variables in the model. In the hierarchical topic model, these latent variables provide the tree structure and node parameters. As exact inference is intractable, we resort to Gibbs sampling for approximate inference.

We modify the Gibbs sampling method in hLDA to estimate the posterior for RShLDA. The core idea of Gibbs sampling is to sample from a Markov chain whose stationary distribution is the posterior. Each latent variable is iteratively sampled conditioned on the observations and all the other latent variable. Following [7]'s approach, we employ collapsed Gibbs sampling, in which we marginalize out the topic parameters β and per-document topic proportions θ_d to speed up the convergence. Therefore, the posterior we need to approximate is $p(\mathbf{c}_{1:D}, \mathbf{z}_{1:D} | \gamma, m, \pi, \eta, \mu, \tau, \mathbf{w}_{1:D}, \mathbf{y}_{1:D,1:D})$, where γ and η are the hyperparameters of nCRP and the topic-word distribution, $\{m, \pi\}$ is the sticking-breaking parameter for topic proportions. In this work we do not provide the parameter learning algorithm and these parameters are fixed according to the analysis and prior expectation about the data, which is discussed in the section VI. In each iteration, the Gibbs sampler is divided into two steps: 1) For each document d , re-sample the per-word level allocations $z_{d,n}$ and per-document paths \mathbf{c}_d based on

observations of \mathbf{w}_d ³, and 2) For each pair of documents (d, d') , re-sample the per-word level allocations $z_{d,n}$ and per-document paths \mathbf{c}_d based on observations of $y_{d,d'}$.

1) *Gibbs sampling based on $\mathbf{w}_{1:D}$* : The first step includes two parts: re-sample the per-word level allocations $z_{d,n}$ with given path assignments, and re-sample the per-document paths \mathbf{c}_d with given level allocations.

Sampling Level Allocations. Given the current path assignments, we need to re-sample the level allocation variable $z_{d,n}$ for word n in each document d :

$$p(z_{d,n} | \mathbf{z}_{-(d,n)}, \mathbf{c}, \mathbf{w}, m, \pi, \eta) \propto p(w_{d,n} | \mathbf{z}, \mathbf{c}, \mathbf{w}_{-(d,n)}, \eta) p(z_{d,n} | \mathbf{z}_{d,-n}, m, \pi) \quad (5)$$

This is the same notation with the standard hLDA [7]. ShLDA differs with it in the first term, which is the probability of a given word based on a possible assignment. In standard hLDA, it is assumed that the topic parameters β are generated from a symmetric Dirichlet distribution:

$$p(w_{d,n} | \mathbf{z}, \mathbf{c}, \mathbf{w}_{-(d,n)}, \eta) \propto \#[\mathbf{z}_{-(d,n)} = z_{d,n}, \mathbf{c}_{z_{d,n}} = \mathbf{c}_{d,z_{d,n}}, \mathbf{w}_{-(d,n)} = w_{d,n}] + \eta \quad (6)$$

where $\#[\cdot]$ counts the elements of an array satisfying a given condition. Let $q_{d,n}$ denote the left side of the above equation. We incorporate the supervision of seed word set by setting a soft constraint to modify the Gibbs sampling process that seed words tend to be generated from the root topic ($z_{d,n} = 1$):

$$\hat{p}(w_{d,n} | \mathbf{z}, \mathbf{c}, \mathbf{w}_{-(d,n)}, \eta) \propto q_{d,n} \cdot \text{Constraint}(\mu, z_{d,n}) \quad (7)$$

where the definition of $\text{Constraint}(\mu, z_{d,n})$ is in Eq.3. Following this sampling process, the words relevant to the query are guaranteed to have a higher probability to be assigned the root topic, leaving the subtopics focusing more on refined terms. The second term in Eq.5 is a distribution over levels which is concerned with the GEM distribution of the stick breaking process. We keep it unchanged.

The step of *sampling path assignments* is identical to that in standard hLDA (see [7] for detail).

2) *Gibbs sampling based on $\mathbf{y}_{1:D,1:D}$* : Within each iteration, after re-sampling the per-word level allocations $z_{d,n}$ and per-document paths \mathbf{c}_d based on observations of \mathbf{w}_d , the second step is to modify the posterior probabilities based on the observed duplicate links $\mathbf{y}_{1:D,1:D}$.

For each pair of documents (d, d') , if the binary duplicate link $y_{d,d'} = 0$, we can not determine the topic similarity of the two documents and thus keep the posterior probabilities unchanged in this iteration. For path assignments \mathbf{c}_d and $\mathbf{c}_{d'}$, if $y_{d,d'} = 1$, the probabilities are updated as:

$$p(\mathbf{c}_d | \mathbf{c}_{-d}, \mathbf{z}, \eta, \gamma, \mathbf{w}_d, \mathbf{y}_{d,d'}) = 0.5 * [p(\mathbf{c}_d | \mathbf{c}_{-d}, \mathbf{z}, \eta, \gamma, \mathbf{w}_d) + p(\mathbf{c}_{d'} | \mathbf{c}_{-d'}, \mathbf{z}, \eta, \gamma, \mathbf{w}_{d'})] \quad (8)$$

$$p(\mathbf{c}_{d'} | \mathbf{c}_{-d'}, \mathbf{z}, \eta, \gamma, \mathbf{w}_{d'}, \mathbf{y}_{d,d'}) = 0.5 * [p(\mathbf{c}_d | \mathbf{c}_{-d}, \mathbf{z}, \eta, \gamma, \mathbf{w}_d) + p(\mathbf{c}_{d'} | \mathbf{c}_{-d'}, \mathbf{z}, \eta, \gamma, \mathbf{w}_{d'})] \quad (9)$$

where on the right side is the posterior probabilities computed from the first step. In this way, we restrict the path assignments of document d and d' to be the same if their duplicate link $y_{d,d'} = 1$, i.e., the subtopics of the documents are identical if they are duplicate with each other.

For level allocations $\mathbf{z}_{d,1:N_d}$ and $\mathbf{z}_{d',1:N_{d'}}$, if $y_{d,d'} = 1$, the probabilities are updated as:

$$p(z_{d,n} | \mathbf{z}_{-(d,n)}, \mathbf{c}_d, m, \pi, \eta, \tau, \mathbf{w}_d, \mathbf{y}_{d,d'}) \propto p(y_{d,d'} | \mathbf{c}_d, \mathbf{c}_{d'}, \theta_d, \theta_{d'}, \tau) \cdot p(z_{d,n} | \mathbf{z}_{-(d,n)}, \mathbf{c}_d, m, \pi, \eta, \mathbf{w}_d) \quad (10)$$

where the first term on the right side is computed according to Eq.4, θ_d and $\theta_{d'}$ are computed by the level assignments $\mathbf{z}_{d,1:N_d}$ and $\mathbf{z}_{d',1:N_{d'}}$ from the step, while the second term on the right side is the posterior probabilities from the first step.

With these conditional distributions, the full Gibbs sampling process is specified. Given current state of the sampler, $\{\mathbf{c}_{1:D}^{(t)}, \mathbf{z}_{d,1:D}^{(t)}\}$, we iteratively sample each variable conditioned on the rest:

- i. For each document d :
 - (a) Randomly draw $\mathbf{c}_d^{(t+1)}$ as in standard hLDA.
 - (b) Randomly draw $\mathbf{z}_{d,1:N_d}^{(t)}$ from Eq.5.
- ii. For each pair of documents d, d' :
 - (a) Update $\mathbf{c}_d^{(t+1)}$ and $\mathbf{c}_{d'}^{(t+1)}$ from Eq.8 and Eq.9.
 - (b) Update $\mathbf{z}_{d,1:N_d}^{(t)}$ and $\mathbf{z}_{d',1:N_{d'}}^{(t)}$ from Eq.10.

After running for sufficiently iterations, we can approach its stationary distribution, which is the conditional distribution of the latent variables in the RShLDA model given the corpus, seed word set and duplicate links.

V. QUERY EXPANSION AND DUPLICATE DETECTION

A. Query Expansion

Query expansion (QE) is the process of reformulating a seed query to improve retrieval performance in information retrieval operations. In our case, we employ query expansion, combining WordNet and association mining to extend the query terms into a seed word set \mathcal{S} , which guide the discovery of the root topic in the derived topic hierarchy.

WordNet [19] is an online lexical dictionary which describes word relationships in three dimensions of Hypernym, Hyponym and Synonym. It is organized conceptually. According to our mechanism of incorporating the supervision information, adding noisy words not included in the vocabulary will not detract from the topic modeling process. This means we are allowed to extend the query as much as we can, on condition that no words concerned with subtopics are mixed. Therefore, we exclude words having hyponym or troponym relations to the query in

³ The first step is the Gibbs sampling for the proposed ShLDA.

WordNet. In addition, instead of removing unusual words, we employ association mining and add high-frequency words into the seed word set.

We utilize WordNet as the basic rule to extend the query along two dimensions including hypernym and synonym relations. Since WordNet has narrow coverage for domain specific queries [20], we use association rules to exploit collection-dependent word relationships. We examine the vocabulary and add the words with both top 10 highest *confidence* and *support* with the original query words into the query expansion. For example, the final seed word set of query ‘9/11 attack’ is $S = \{911\ attack\ Assault\ aggress\ assail\ fight\ struggle\ contend\ onslaught\ onset\ attempt\ operation\ approach\ event\ wtc\ world\ trade\ center\ terrorist\ terrorism\ 9-11\}$.

B. Duplicate Detection

Near-duplicate provides strong cues for linking related videos. If two videos are near-duplicate, they can be regarded as having a must-link constraint, which indicates that they describe the same subtopic and should be grouped into the same cluster. A binary variable $y_{d,d'}$ is employed to denote the duplicate link between each two videos d and d' , which is then formulated as the observed pairwise link and integrated into the topic modeling process (see section IV.C).

We utilize the algorithm proposed in [21] for duplicate detection, which is state-of-the-art method. Sampled frames are matched based on a novel local feature indexing method, which is robust to video transformations and efficient in memory usage and computation time. The spatio-temporal consistency are verified after individual frames matching and the score for video-level matching are normalized. The technical details can be referred to the paper.

The total number of videos in our collected dataset is 5,405. We evaluate the performance of duplicate detection in the experiments. The number of duplicate videos is 689 and the duplicate rate is 12.7%, which is lower than the demonstrated number [16] probably because we only analyze the top returned videos. The F-measure of duplicate detection is 75.4%. We emphasize that duplicate detection can be performed when the videos are uploaded, reducing the online time cost, which is very important for search result clustering.

VI. EXPERIMENTS

A. Dataset and Evaluation Metrics

Since the goal of this paper is to present a multi-modal framework for clustering-based web video retrieval, we assess the retrieval effectiveness in a web-scale video dataset collected from video sharing websites. After careful examination of the hottest topics in Youtube, Google Zeitgeist, and Twitter, we selected seven social and political topics as queries. We issued these queries to Youtube, Metacafe and Vimeo, and crawl the top 500, 150 and 150 (if there are) returned videos for experiments,

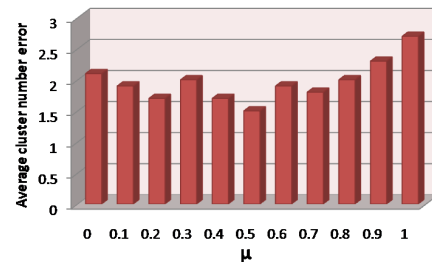


Figure 4. Average subtopic number error as μ changes.

respectively. We focused on the topmost search results to avoid bringing too many unrelated videos. Videos with no tags are filtered out. The videos collected from each query form a video set. The queries and statistics about the corresponding video set are listed in Table I.

To evaluate the performance of proposed methods, we implement BCS [5], LDA and the standard hLDA as the baseline for comparisons. As far as we know, BCS is the only existing work addressing the problem of video search result clustering. They cluster the top returned videos based on visual similarity of low-level appearance features and textual similarity of term vector features. We compare the different methods on retrieval performance and the clustering quality. For the retrieval performance, we use the metrics of *subtopic reach time* (SRT, [22]) and *time cost*. SRT is proposed in text search result clustering, which is a modelization of the time taken to locate a relevant document. For the clustering quality, we access the metrics of *purity* [23], *F measure* [24] and *clustering description readability*.

B. Parameter Settings

Topic models make assumptions about the topic structure by the settings of hyperparameters. We empirically fixed the hyperparameters according to the prior expectation about the data. The hyperparameter η controls the smoothing/sparsity of topic-word distribution. Small η encourages more words to have high probability in each topic. (For LDA, it requires less topics to explain the data. For hLDA, ShLDA and RShLDA, it leads to a small tree with compact topics.) Delighted from this, we empirically chose a relatively small value of η and set $\eta = 0.5$. All the hierarchical topic models have an additional hyperparameter, CRF parameter γ , which decides the size of the inferred tree. As in [25], we set $\gamma = 1$ to reduce the likelihood choosing new paths when traversing nCRP.

We then discuss the choice of supervision strength parameter μ in ShLDA and the coefficient parameter τ in RShLDA. We utilize a benchmark text subtopic retrieval dataset, AMBIENT⁴, for the determination of μ . We assume that appropriate μ brings no perturbation to the hierarchical topic discovery process and the derived topic tree should be consistent with the latent hierarchical structure. Therefore, we analyzed the error between the

⁴ <http://credo.fub.it/ambient>

TABLE I.
COLLECTED WEB VIDEO DATASET STATISTICS

ID	Query	#Video retrieved	#Video collected	#Vocabulary	#Words
1	9/11 attack	8,361	791	2140	38747
2	gay rights	602,885	799	2048	35538
3	abortion	66,606	797	1770	33144
4	Iraq war invasion	4,425	702	1778	36760
5	Beijing Olympics	202,511	787	1718	32370
6	Israel palestine conflict	252,746	798	1814	38499
7	US president election	36,037	731	1792	33249

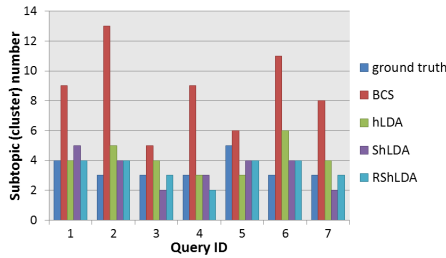


Figure 5. The ground-truth subtopic number and automatically derived cluster number for the test queries

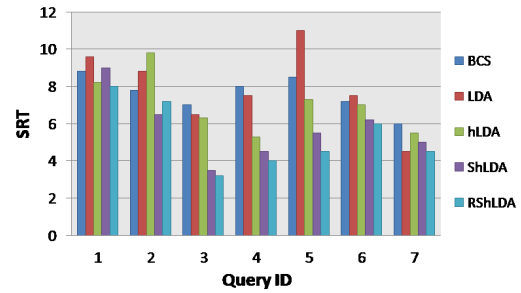


Figure 6. SRT of different clustering methods

subtopic number of ground truth and the derived subtopic number of ShLDA over the different values of μ (see Fig.4). $\mu = 0.5$ achieves the least error. Therefore, we fixed $\mu = 0.5$ in following experiments. As we build a two-level topic hierarchy and all the videos share the same root topic, τ is a two-dimensional vector and τ_1 can be set to 0. τ_2 is empirically set as 0.1.

The most important parameters for BCS are the weights for adopted features, visual, tag, title and description. Affinity propagation (AP) and normalized cut (NC) are utilized as the clustering algorithm and they demonstrated AP generally outperforms NC. Therefore, we fixed the set of feature weights showing best performance with AP clustering: visual-0.3, tag-0.49, title-0.07, description-0.14.

Furthermore, we notice that, AP-based BCS, hLDA and the proposed ShLDA and RShLDA can automatically determine the number of clusters, which is reasonable in practical implications. However, LDA must be given the number of clusters as input. Choosing the appropriate number of clusters is challenging, especially for post-retrieval clustering task. To simplify the experimental configurations, we fix the cluster number as the ground truth value for LDA.

C. Experimental Results

1) *Visualization of the discovered subtopics:* We visualize the discovered subtopics of video collections for the seven test queries in Fig.11 and Fig.12. For the query of '9/11 attack', the subtopics derived from LDA and topic hierarchies derived from hLDA, ShLDA and RShLDA are presented in Fig.11 for comparison. It is shown that LDA mixes common words like 'attack', '9/11', 'September', 'terrorist' in different subtopics and fails to discover the shared topic. Compared with flat structure based clustering method of LDA, utilizing the hierarchical topic models will prevent the shared topic

from being mixed within other topics and thus ensure the clustering performance. The topic hierarchy recovered by hLDA finds the shared topic on the root level. However, without constraint of topic distribution over the seed word set, words describing the shared topic, e.g. 'wtc', 'terrorist', '11', 'attack' also appear in subtopics. This contaminates the subtopics and limits its power to sub-events or viewpoints detection. Incorporated with supervision information, ShLDA prevents seed words generating from the subtopics, and results in a topic hierarchy with subtopics focusing on the refined themes. RShLDA further improves over ShLDA by considering the visual cues. With the aid of duplicate constraints, the visual similar videos are guaranteed to be clustered together and RShLDA uses link information to influence the topic hierarchy and the subtopic distribution.

2) *Comparing different clustering methods:* For evaluation, human accessors create ground-truth subtopic themes after browsing the retrieved videos for each query-corresponding video set. For example, the subtopic themes inside the video collection derived from the query 'abortion' are summarized as pro-abortion, anti-abortion and neutral. Videos are manually labeled as belonging to a certain subtopic (cluster). The ground-truth subtopic number and derived subtopic (cluster) number by BCS, hLDA, ShLDA and RShLDA for the test queries are shown in Fig.5. We can see that all four models fail to recover the ground-truth subtopic number for some video sets. The reason is that the ground-truth subtopic themes created by subjective assessment may not reflect the nature of the video set, especially when unrelated noisy videos are involved. We also notice that RShLDA, ShLDA and hLDA performs better than BCS. The BCS curve is high above the ground truth. This is due to its duplicate clustering alike mechanism, which results in small-size duplicate video clusters.

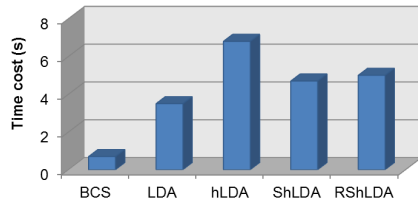


Figure 7. Average time cost

We first compare the SRT and time cost of different clustering methods (see Fig.6 and Fig.7). Note that BCS and RShLDA utilize additional visual information. From Fig.6, we can see the best performance is achieved by RShLDA, followed by ShLDA and hLDA, which is due to the multi-modal framework and the separation of shared common topic from subtopics. It is interesting to see that the BCS basically outperforms LDA, and achieves even better SRT than ShLDA for query 1. The reason is that BCS additionally utilizes visual information for clustering. However, BCS suffers from two problems: 1) it employs a flat-structure clustering algorithm; 2) it uses the cluster centroid to represent the cluster and provides no mechanism for how to derive the cluster labels.

For clustering-based video retrieval, the clustering is performed online, which requires necessarily short response time. We focus on the efficiency of clustering algorithms and do not consider about the video acquisition time cost. We assume that visual features used in BCS are extracted offline and take no account of text preprocessing time. Fig.7 illustrates time complexity for the clustering algorithms. For ShLDA and RShLDA we do not consider the query expansion time, and the query expansion time from local-stored WordNet is about 1.4s. As duplicate detection can be performed when videos are uploaded, we also exclude the time cost of duplicate detection for RShLDA. Since BCS uses AP for clustering, it achieves lower time cost than the generative topic models. The speedup of ShLDA over hLDA is due to that incorporated prior guides the seed words gradually generated from the root set and thus speeds up the convergence process. RShLDA costs a little more training time than ShLDA as the available positive duplicate links are very small compared to the huge pairwise link number.

We noticed that the computational cost dramatically increases when dealing with large-scale web videos, and we will be researching towards this in future work. In addition to develop faster hierarchical topic models, we can design a composite UI to improve user experience while waiting for the cluster-based results: The search engine returns the conventional list-based results as soon as users issue the queries. The server conducts topic modeling while users browse through the list-based results. Once the training is finished, the videos will be mapped to the corresponding subtopics and cluster-based results will be presented to the users. Instead of online training, we can also perform offline training for all pre-chosen queries on the server. Two schemes are considered for offline training. The first is re-training models for all the queries at intervals on the server, when bunches of new videos are uploaded.

The second is estimating the topic (related queries) of the uploaded videos, and developing algorithms for model update and modifying topic models for related queries.

Besides accessing the retrieval performance, we also evaluate the clustering qualities. Cluster purity is used to evaluate the performance of video clustering:

$$purity = \frac{1}{N} \sum_{i=1}^N \frac{\max_j |C_i \cap S_j|}{|C_i|} \quad (11)$$

where N is the number of videos in the collection, C_i is the set of videos in the i^{th} cluster and S_j is the set of videos in the j^{th} subtopic. Fig.8 shows the cluster purity. We find that BCS noticeably outperforms the other algorithms. High purity is easy to achieve when the number of clusters is large. Therefore, we cannot use purity to trade off the quality of the clustering against the number of the clusters. A measure to make this tradeoff is *F measure*. We evenly penalize false negatives and false positives, i.e. the F1 measure (see Fig.9). It is shown that BCS performs poorly on F1 measure, even much worse than LDA. The reason is that BCS focuses on clustering duplicate or near-duplicate videos, which limits the cluster size and forces considerable number of semantically similar videos assigned to different clusters.

The quality of the cluster description is crucial to the usability of clustering-based video retrieval. If a cluster cannot be described, it is presumably of no value to the user. BCS employs the cluster centroid as the cluster representation, which lacks real descriptions and is of little use for guiding the user understanding the cluster content. The cluster description readability is evaluated as follows. Each subtopic characterized by the top five probable words was shown to the participants with the top three ranked videos within this subtopic. The participants were asked to evaluate the cluster description readability in two aspects: ‘whether the topic description itself is sensible, comprehensive and compact’ (question 1) and ‘whether the topic description is consistent with the representative videos’ (question 2). For each question, participants rated from one to five where five is best. The average ratings are shown in Fig.10. The proposed RShLDA and ShLDA show superiority on generating meaningful cluster descriptions, especially on generating sensible, comprehensive and compact representations (question 1). We note that ratings for query 5 - ‘Beijing Olympics’ are relatively low. In the retrieved video set of ‘Beijing Olympics’, diverse events or subtopics are involved, e.g. opening ceremony, game video, athlete interview, torch relay, etc. The discovered topic structure is sparse and less meaningful. Besides, some unrelated videos regarding issue of Tibet are also included.

VII. CONCLUSIONS

In this paper, we have presented a hierarchical topic model based multi-modal framework for web video faceted subtopic retrieval. Instead of showing a long ranked list videos, we explore the hierarchical topic

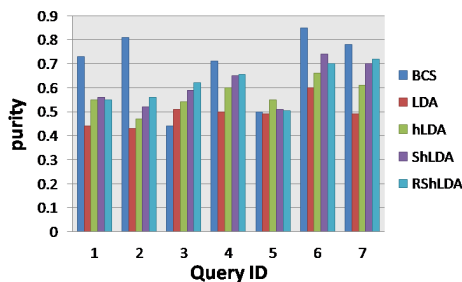


Figure 8. Purity rates

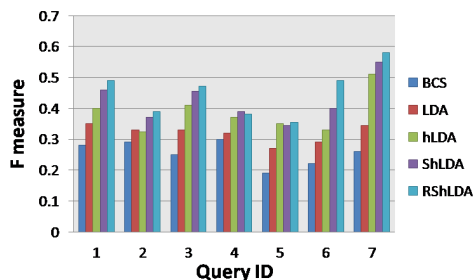


Figure 9. F measure for different methods.

structure in the retrieved video collection returned from faceted queries and present users with videos organized by semantic clusters. Experiments demonstrate the effectiveness of the proposed method.

In the future, we will improve our current work along the following directions. 1) Unrelated videos in retrieved video collections will affect the clustering performance. We will develop noisy subtopic aware hierarchical topic model to reduce the influence of noises as well as remove unrelated videos. 2) In this paper, we do not provide the algorithm of learning the hyper parameters. Posterior inference and parameter estimation are two basic computation issues for nonparameteric Bayes modeling. Designing the parameter estimation algorithm is one of our future work.

VIII. ACKNOWLEDGEMENT

This work was in part supported by National Natural Science Foundation of China (Grant No. 90920303) and National Program on Key Basic Basic Research Project (the 973 Program, Project No. 2012CB316304).

REFERENCES

- [1] S. Osinski and D. Weiss, "A concept-driven algorithm for clustering search results," *Intelligent Systems, IEEE [see also IEEE Intelligent Systems and Their Applications]*, vol. 20, no. 3, pp. 48–54, 2005.
- [2] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma, "Learning to cluster web search results," in *SIGIR*, 2004, pp. 210–217.
- [3] D. Cai, X. He, Z. Li, W. Y. Ma, and J. R. Wen, "Hierarchical clustering of www image search results using visual textual and link information," in *ACM Multimedia*, 2004, pp. 952–959.
- [4] F. Jing, C. Wang, Y. Yao, K. Deng, L. Zhang, and W. Y. Ma, "Igroup: web image search results clustering," in *ACM Multimedia*, 2006, pp. 377–384.
- [5] A. Hindle, J. Shao, D. Lin, J. Lu, and R. Zhang, "Clustering web video search results based on integration of multiple features," *World Wide Web*, vol. 14, no. 1, pp. 1–21, 2010.
- [6] Z. M. Zhang, Z. Zhang, R. Jain, and Y. Zhuang, "Overview of acm international workshop on connected multimedia," in *Proceedings of the international conference on Multimedia*, 2010, pp. 1763–1764.
- [7] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies," *J. ACM*, vol. 57, no. 2, pp. 1–30, 2010.
- [8] B. Carterette and P. Chandar, "Probabilistic models of ranking novel documents for faceted topic retrieval," in *Proceeding of the 18th ACM conference on Information and knowledge management*, 2009, pp. 1287–1296.
- [9] C. L. Clarke, N. Craswell, and I. Soboroff, "Overview of the trec 2009 web track," Tech. Rep.
- [10] W.-H. Lin and A. G. Hauptmann, "News video classification using svm-based multimodal classifiers and combination strategies," in *ACM Multimedia*, 2002, pp. 323–326.
- [11] L. Liu, Y. Rui, L.-F. Sun, B. Yang, J. Zhang, and S.-Q. Yang, "Topic mining on web-shared videos," in *ICASSP*, 2008, pp. 2145–2148.
- [12] C. Ramachandran, R. Malik, X. Jin, J. Gao, and J. Han, "Videomule: a consensus learning approach to multi-label classification from noisy user-generated videos," in *MM*, 2009.
- [13] J.-Y. Pan, H.-J. Yang, and C. Faloutsos, "Mmss: Multimodal story-oriented video summarization," in *ICDM*, 2004, pp. 491–494.
- [14] D. qing Zhang, C. yung Lin, S. fu Chang, and J. R. Smith, "Semantic video clustering across sources using bipartite spectral clustering," in *Spectral Clustering, International Conference on Multimedia and Expo*, 2004, pp. 117–120.
- [15] M. Guillaumin, J. J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *CVPR*, 2010, pp. 902–909.
- [16] X. Wu, A. G. Hauptmann, and C.-W. Ngo, "Practical elimination of near-duplicates from web video search," in *ACM MM*, 2007, pp. 218–227.
- [17] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 7, pp. 993–1022, 2003.
- [18] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.
- [19] G. A. Miller, R. Beckwith, C. Felbaum, D. Gross, and K. Miller, *Introduction to WordNet: An On-line Lexical Database*. Oxford Univ Press, 1990, vol. 3, no. 4.
- [20] K. Chandramouli, T. Kliegr, J. Nemrava, V. Svatek, and E. Izquierdo, "Query refinement and user relevance feedback for contextualized image retrieval," in *Visual Information Engineering*, Xian China, 2008, pp. 452–458.
- [21] M. Douze, H. Jegou, and C. Schmid, "An image-based approach to video copy detection with spatio-temporal post-filtering," *IEEE Transactions on Multimedia*, vol. 12, no. 4, pp. 257–266, 2010.
- [22] C. Carpineto, S. Osinski, G. Romano, and D. Weiss, "A survey of web clustering engines," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–38, 2009.
- [23] P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison Wesley, 2005, vol. 19, no. 1.
- [24] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *ACM SIGKDD*, 2000, pp. 35–42.
- [25] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. Tenenbaum, "Hierarchical topic models and the nested chinese restaurant process," in *Advances in Neural Information Processing Systems*. MIT Press, 2004, pp. 17–24.

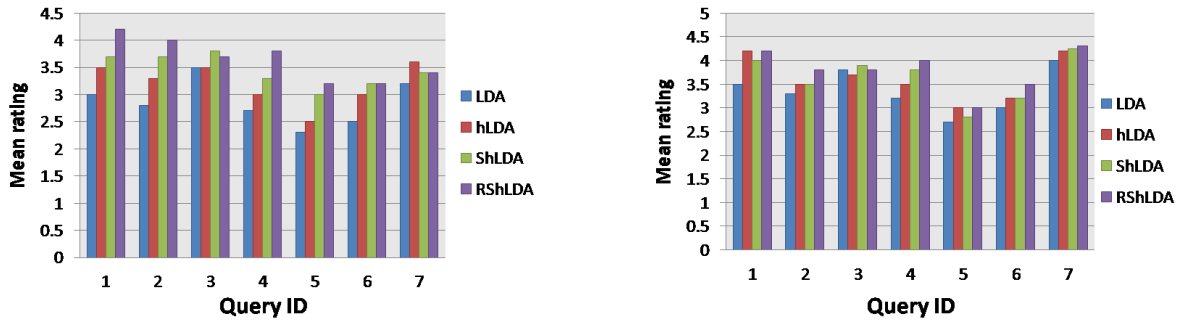


Figure 10. Mean ratings of cluster description readability for (left:) Question 1 (right:) Question 2.

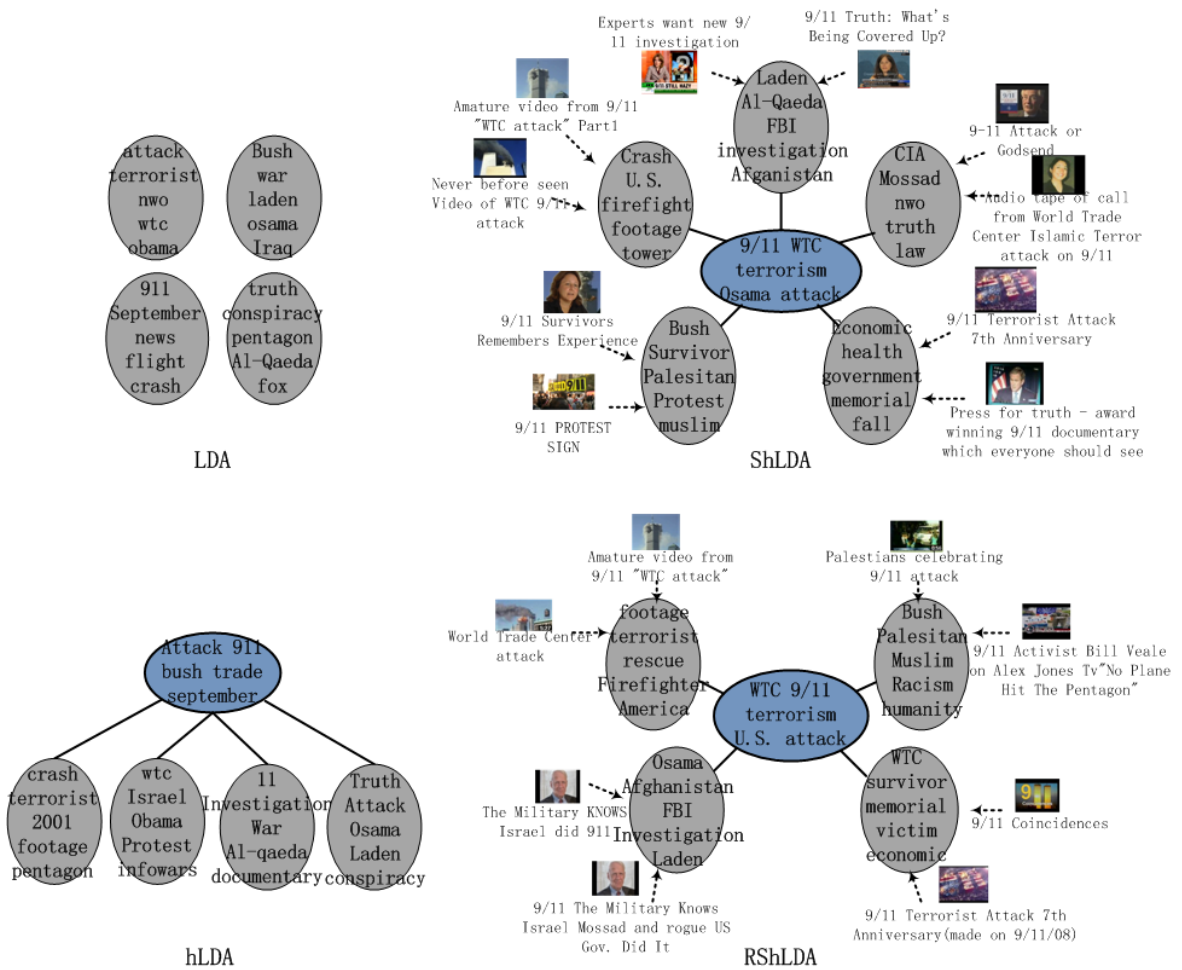


Figure 11. Discovered topic hierarchy from the video collection of query of '9/11 attack' using LDA, hLDA, ShLDA and RShLDA. For ShLDA and RShLDA, we also present two videos having the largest proportion associated with the subtopics

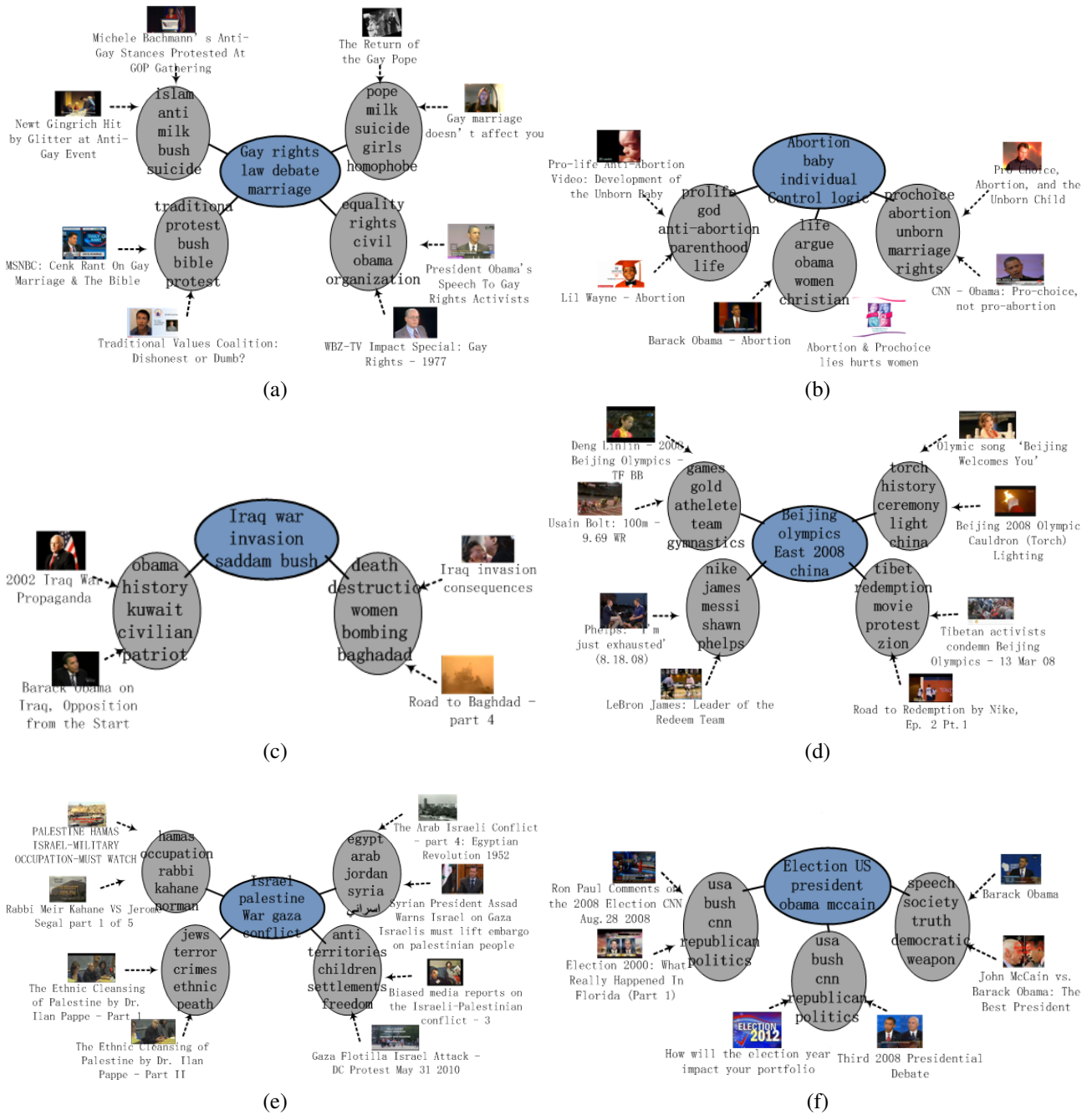


Figure 12. Discovered topic hierarchies from the video collection of queries from Youtube using RShLDA. (a) 'gay rights' (b) 'abortion' (c) 'Iraq war invasion' (d) 'Beijing Olympics' (e) 'Israeli Palestine conflicts' (f) 'US president election'.

Tour Route Recommendation Begins with Multimodal Classification

Xiujun Chen

School of Computer Science and Engineering
Northwestern Polytechnical University, Xi'an 710072, China
Email: xiujun_chen@hotmail.com

Qing Wang

School of Computer Science and Engineering
Northwestern Polytechnical University, Xi'an 710072, China
Email: qwang@nwpu.edu.cn

Abstract—Location estimation of tourist photos by classification is challenging due to the unstable and less-discriminative features of photos taken in each city. In this paper, we originally deal with this issue in an alternative way, which begins with but is not limited to classification. We explore textual, temporal, geographic as well as visual information to make tour route recommendation. Given a query photo, we recommend a city by first classifying the photo to get scores that indicate its similarities with photos from each city, and then evaluating the attractiveness of each city by modeling its hotness potential. We do not aim at finding out where the photo was taken exactly; instead, we combine the similarity and hotness potentials according to the user's preference and recommend the city that has the highest combination value. Then, we suggest the next city by further modeling the correlation and interaction between cities pairwise with two potentials, i.e., proximity and co-visitedness. Applying the greedy algorithm, we recommend one city at a time and eventually generate a tour route consisting of all the recommended cities in order. Furthermore, in order to show different visual and cultural characteristics across cities, we classify photos of each city into four categories, i.e., food, landscape, man-made and person. In experiments, we collected a database containing 41792 photos of 35 important cities along the silkroad and provided a query sample for tour route recommendation. Experimental results have shown the effectiveness and reliability of our recommendation model.

Index Terms—tour route recommendation, image classification, city hotness, photo location estimation, silkroad

I. INTRODUCTION

In everyday life, we come across numerous photos from time to time. Also, we are frequently amazed by some of the wonderful photos that grasp our attention at first glance. We may begin to wonder where these photos were taken. Suppose someone is going to travel in the summer vacation, and he/she believes that the location with sights in a particular photo is the right place to go. But where is it? This question is extremely difficult to answer. Let's take a look at photos in Fig. 1 first.



Figure 1. I would like to travel there! Could you please tell me where they are located?

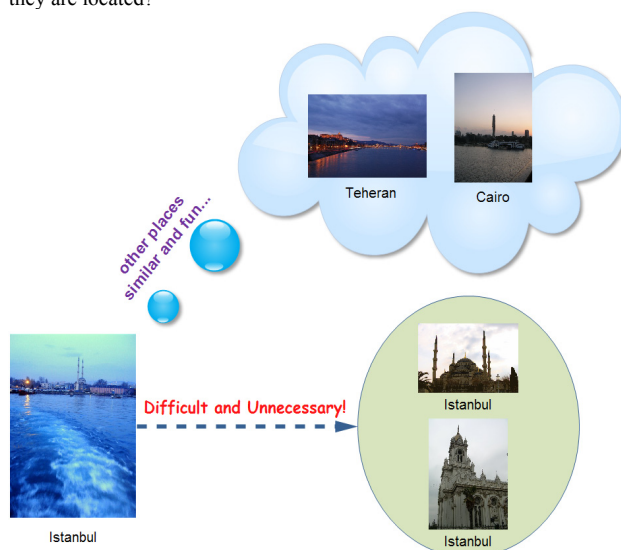


Figure 2. The illustration of our motivation. The query photo was taken in Istanbul, but its content is overwhelmed by sea, which is not representative or typical for a city. Therefore, we can hardly relate this photo to the others which were also taken in Istanbul, such as the two photos within the circle in the lower right corner. As an alternative, we propose a recommendation model consisting of four potentials to find out other interesting place which has sights similar to the query photo. In this example, we recommend Teheran and Cairo, which also possess photos with a large area of sea or water.

These three photos are taken in Cairo, Xian, and Istanbul respectively. Although they are all famous sights for traveling, it is still not easy to identify them if we have never visited them before.

Multiple approaches have been proposed to address the problem of photo location estimation by classification. Li

et al. [2] classified images with textual and temporal information as well as visual features and achieved a competitive result when considering photos as streams. Our work is similar to Li's method [2] in the belief that photos are not completely separated. Instead, they should be taken in a stream, from which we can explore the correlation between neighbor photos and analyze the information they convey mutually. However, the dataset they used contains as many as 30 million images, which makes it difficult to generalize to other applications. Hays et al. [1] estimated the relatively accurate geographic information of an image in a purely data-driven way using various features such as color, GIST and texon. But the GPS tags they used are difficult to obtain for most of the time. To deal with the issue of dataset, Guillaumin [3] proposed a semi-supervised method based on MKL (multiple kernel learning) to automatically label the unlabeled data, and thus enlarging the training dataset to learn an ultimate visual classifier. Unlike this semi-supervised framework, we adopt a method proposed by Wang [4], which build textual features for testing image straightforwardly by adding up features of its nearest neighborhoods in visual feature space.

Different from the methods mentioned above, which aim at increasing the classification accuracy, we do not make city recommendation purely depending on classification results, which is still involved with currently unsolved problems. Our system is based on the assumption that people who are interested in the photos are not necessarily looking for the exact places where they were taken. Instead, people are more likely to be interested in the style of the scenes in the photos. Therefore, other places, even the "wrong" places but share similar sights and culture with the target city are acceptable and may be even more attractive. The motivation of this paper is illustrated in Fig. 2.

On this assumption, we propose a novel framework that makes recommendations associated with some priors, which are modeled by so-called potentials. We consider the classification scores as only one of our four potentials: similarity potential. We call it similarity since it indicates the similarity between the query photo and example photos taken from each city. Besides, we model the hotness potential that indicates how frequent the city is visited. We combine the hotness potential with the similarity potential to recommend the first city to user. Furthermore, we aim to generate a complete tour route rather than simply one-city query. Thus, we model the proximity potential to prefer the next recommended city to be less distant. Moreover, the co-visitedness potential, which is modeled by exploring and following the example of other tourists, is also introduced to make the system incline to the city that is commonly co-visited with the current city in the tour route. In summary, we combine these four potential to make city recommendation for each query (except the first query). Note that the similarity potential is obtained from classification results and the other three potentials are mined from information contained in photos. A brief illustration of potential models for recommendation is shown in Fig. 3.

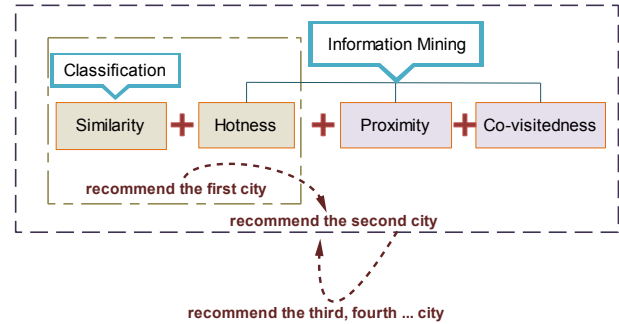


Figure 3. Brief illustration of potential models for recommendation. The similarity potential is obtained by classification and the other three potentials are mined from the information of photos. We recommend the first city with similarity and hotness potentials, and subsequent cities with all the four potentials.

In addition, we classify the photos of each city into four categories, i.e. food, landscape, man-made and person, to show the different characteristics of cities and help users to find attractive photos easier. The pipeline of our recommendation framework is described in detail in Section IV with Fig. 5.

From the viewpoint of information retrieval, our work can be considered as a cross-media retrieval system [17], [18], [19], which retrieves the location of the photo-taken place by multiple modalities such as images and tags. Rather than follow the general model of MMD semantic graph [17], [18], [19], we specialize our work to deal with the tour city problem so that we can explore the specific relationship between different modalities and fully exploit the information conveyed by them. In fact, the principle of our co-visitedness model is similar to the PageRank algorithm [6], which mines the link relationships among objects to benefit the retrieval. Note that the goal of our system is different from that of traditional retrieval since we do not aim at finding out the exact location of the photo-taken place. Instead, we would like to combine multimodal information to make tour recommendations according to users' preference.

The rest of paper is organized as follows. In Section II, we present the classification method that we adopt to calculate the similarity potential. In Section III, we introduce the other three potentials we mined from the information contained in photos. In Section IV, we present our final recommendation model by combining the four potentials and summarize the pipeline of our framework. Section V shows the experimental results and provides a sample of query for cities along the silkroad. Finally we draw conclusions and propose possible future work in Section VI.

II. SIMILARITY POTENTIAL FROM CLASSIFICATION

Since the scores of classification indicate the similarity between the query photo and the photos in each city collection, we model the similarity potential by simply normalizing the classification scores. To perform this classification, we combine the visual and textual features of each query photo to build the final features and then feed them into the offline trained classifier to obtain the scores. Below is the introduction of visual features, textual features and the combination of them.

Visual features: We first extract SIFT descriptors [7] in a dense grid. Then we follow the work in [12], [13], which both consider vector quantization and sparse coding as coding strategies, and max pooling and average pooling as pooling strategies. For vector quantization, we use k-means to cluster a dictionary of size 2000 and then quantize each of the SIFT vectors into words from the learned dictionary. For sparse coding, which is suggested in [15] as a generalization of VQ, we also learn a dictionary of size 2000 and then ensure sparsity when encoding SIFT descriptors. After that, the encoded image can be represented as a $P \times 2000$ matrix, where P is the number of the SIFT features. By pooling process, we can reduce each column of the matrix into a scalar, and thus reduce the whole matrix into a single 2000 dimension vector. In this paper, we evaluate two pooling strategies, i.e., average pooling that accumulates all the components in the column, and max pooling that picks out the max component in the column. We will compare the efficiency of these strategies and find out the best one to accomplish our classification work.

Textual features: Unlike the setup of some traditional image classification, our training photos downloaded from Flickr [5] are associated with textual features. For the textual features, we build a tag dictionary of size 1030, which contains tag words that appear in frequency higher than 26 times. Therefore, we finally obtain a 1030-dimension histogram for each photo as its textual feature.

However, when it comes to the testing phase, this textual information is no longer attached to the test photos. To address this problem, we adopt the method proposed by Wang [4] and use a reference set to estimate the textual features. The reference set consists of photos whose textual features have been pre-computed and associated. In particular, the training set can be used as reference set. For each test photo, we find its 10 nearest neighborhoods from the reference set with visual features. Then we build the textual feature by adding up the features of all the top 10 nearest neighborhoods.

Combination: In this work, we adopt the following two methods of combining the visual and textual features to form the final features.

The first one is to concatenate the two feature vectors. To make the concatenated feature discriminative, we use the model similar to [4] to learn a concatenation weight for these two features. The objective function is defined as follows,

$$J = \sum_i (e^{-d_v - wd_i} - S_i)^2, \quad (1)$$

where i denotes the i^{th} photo pair to train the concatenation weight, d_v and d_i are the distances of visual and textual features for the i^{th} pair respectively. We define $S_i = 1$ when two photos of the i^{th} pair belong to the same category, and $S_i = 0$ otherwise.

By minimizing the objective function J defined in (1), we force the photos to be close in the concatenated feature space if they are from the same category and to be far away if they are not.

The second method is the Multiple Kernel Learning (MKL) that can learn the weights for combining a kernel constructed by the visual features with a second kernel corresponding to the textual features, and learn the parameters of support vector machine (SVM) which we adopt as our classifier simultaneously.

We adopt the classification method that works best in our experiments and normalize the classification scores of this method to obtain the similarity potential.

III. POTENTIALS FROM INFORMATION MINING

Each photo downloaded from Flickr by im2gps code [1] has its comment field, containing information such as photo ID, title, description, date upload, owner, and interestingness. Among them, we extract the following information from each photo, which reveals the relative importance of photos and indicates the correlation and interaction between them.

- **Owner:** it shows the ID of the Flickr user who uploaded the photo. Since the ID is unique for each user, we can use it to identify the users. We extract this information to make hotness inference, which will be introduced later in detail. Although the author and owner of a photo can be different, we do not distinguish them in this paper.

- **Owner's name:** it is the name of the 'Owner', which is more straightforward and readable than the ID. We use owner names when displaying the hottest authors.

- **Date upload:** it is a numerical form of the date when the user uploaded the photo. We extract this information to calculate the time interval between the dates when photos of two cities were uploaded by the same author. We use this temporal information to model the correlation between two cities, which we will introduce in the part C of this section in detail.

- **Interestingness:** it is the rank of all the photos returned by Flickr, which indicates the relative popularity and significance of the photo. For convenience, we transform it into numerals and then normalize it into (0 1).

- **Latitude and Longitude:** these two properties point out the geographic location of photo-taken city. We assign the geographic tag for each city with the most frequent value of latitude and longitude in its photo collection.

By using the above information, we model the following three potentials of each city or city pair for recommendation.

A. Hotness Potential of each city

Generally speaking, tourists tend to take more photos in cities where they find something interesting. In turn, these photos taken by them are likely to attract other tourists to visit the interesting cities. As a result, hotter cities correspond to more photos of higher interestingness and larger numbers of authors. Based on this observation, we model hotness of each city by multiplying the relative interestingness of photos taken there and the relative numbers of authors who have taken photos there. We denote the set of K cities as $C = \{C_k \mid k = 1, 2, \dots, K\}$,

and each city C_k contains a photo collection P_{C_k} . Also, we denote the set of N authors as $A = \{A_r \mid r = 1, 2, \dots, R\}$, and each author A_r corresponds to a photo collection P_{A_r} . Therefore, we get the whole photo set as $P = \bigcup_{k=1}^K P_{C_k} = \bigcup_{r=1}^R P_{A_r}$. Then we compute the interestingness value $I(p_i)$ for each photo p_i . For simplicity, we denote the sum of the interestingness of the photos taken in city C_k as,

$$I_{C_k} = \sum I(P_{C_k}). \quad (2)$$

Now, the hotness of city C_k can be computed as,

$$HC_k = \frac{I_{C_k} \sum_{r=1}^R x_{r,k}}{\sum_{k=1}^K I_{C_k} R}, \quad (3)$$

where $x_{r,k}$ is a binary value, which is assigned 1 if author A_r has taken photos in city C_k , and assigned 0 otherwise. The left factor in (3) is the relative interestingness of the photos in city C_k and the right factor is the relative numbers of authors who have taken photos there.

B. Proximity Potential of each two-city pair

With the latitude and longitude obtained from the photo collection of each city, we can estimate the geographic distance of each city pair. We assume that the earth is an ideal sphere and thus the distance can be easily computed from the latitude and longitude coordinates by applying geometry principles. The ultimate distance between city pair $(C_i - C_j)$ can be modeled as,

$$D_{ji} = D_{ij} = R * \cos^{-1}(\sin g_{lat_i} \sin g_{lat_j} + \cos g_{lat_i} \cos g_{lat_j} \cos(g_{lon_i} - g_{lon_j})), \quad (4)$$

where $R = 6378.1km$ is the equatorial radius of the earth, and g_{lat_i} , g_{lon_i} is the geographic latitude and longitude of city C_i respectively.

Assuming that the users always travel by air and the average speed remains constant (If two cities are too close to take a plane, the length of the actual path of other vehicles, which can be regarded as a curve between two nearby points, will also be close to the length of the direct path), then we can get the geographic proximity of each two-city pair by calculating the reciprocal of distance as,

$$P_{ij} = 1 / D_{ij} \quad (5)$$

Finally, we normalize all the elements of the pair-wise proximity matrix into [0 1].

C. Co-visitedness Potential of each two-city pair

The co-visitedness of each two-city pair evaluates the potential that people who visit one city of the pair will visit the other soon. If two cities are generally visited by the same tourists, it is highly likely that these two cities have an advantage for attracting tourists who have visited one of them. This advantage may result from diverse

scenes for visitors' curiosity, common style for consistent interest, efficient transportation for convenience and so on. Therefore, we can consider the shared tourists as role-models, who are unconsciously followed by other visitors. Empirically, more experienced and active authors are more likely to become role-models. So we model the author hotness first.

Similar to city hotness, the author hotness is modeled based on the observation that hotter author should have taken photos of higher interestingness and have visited larger numbers of cities. For simplicity, we denote the sum of the interestingness of the photos taken by author A_r as,

$$I_{A_r} = \sum I(P_{A_r}). \quad (6)$$

Thus, the hotness of author A_r can be computed as,

$$HA_r = \frac{I_{A_r} \sum_{k=1}^K x_{r,k}}{\sum_{r=1}^R I_{A_r} K}, \quad (7)$$

Now we can simulate the role-model power with author hotness and formulate our primary co-visitedness potential by accumulating the hotness of authors that have visited both cities as,

$$C_{ji} = C_{ij} = \sum_{r=1}^R x_{r,i} x_{r,j} HA_r, \quad (8)$$

where HA_r is the hotness of author A_r modeled by (7), and $x_{r,i}$ is consistent with that in (3), which denotes whether author A_r has taken photos in C_i or not. An HA_r is added when and only when $x_{r,i} = x_{r,j} = 1$, which means that author A_r has visited both the city C_i and city C_j . This is exactly why we call it co-visitedness.

Furthermore, we observe that the role-model power is related to not only the author hotness, but also the time interval between the visiting dates of two cities. In particular, if an author have taken photos of two cities nearly at the same time, it is very likely that he/she has visited both cities in a single travel for some reason. Thus, we prefer to recommend the users to travel these two cities in a single travel as well, by increasing the role-model power of this author for the two-city pair. A similar idea is proposed by Li [2], who states that certain sequences of category labels are much more likely to appear than others, which also indicates the phenomenon of "visit cities by group". Therefore, we should use the visiting time interval to bias the role-model power of authors for each city pair. In general, the shorter the time interval is, the higher the role-model power will be. However, when time interval exceeds a threshold, it will no longer make any differences. For example, if we have visited a city twenty years ago and another city ten years ago, these two cities will have equally little relevance to the city we are visiting now. Therefore, we adopt the logistic function to model the variation of role-model power along with time intervals, as shown in Fig. 4. With this bias of the role-model power of each author who has

visited both cities, we can get the ultimate model of co-visitedness as,

$$C_{ji} = C_{ij} = \sum_{r=1}^A \frac{1}{1 + de^{aT_{ji,r}+b}} HA_r, \quad (9)$$

where $T_{ji,r}$ denotes the normalized time interval between the upload dates of photos taken in city C_i and city C_j by author A_r , and is assigned infinity if A_r has not visited both of the two cities. The parameters d, a, b in this paper is set to 1000, 20, and -10 respectively.

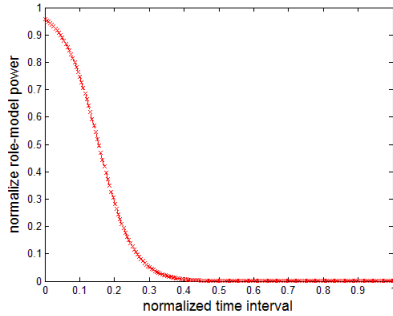


Figure 4. The role-model power of authors corresponding to the time interval between the visits of two cities.

IV. RECOMMENDATION AND THE PIPELINE

We pre-compute all the three potentials discussed in Section III and combine them with the similarity potential estimated by classification scores. Then we can model the final recommendation potential as,

$$E_i = E_{similarity_i} + E_{hotness_i} + E_{proximity_{ji}} + E_{co-visitedness_{ji}} \quad (10)$$

$$= w_1 S_i + w_2 H_{c_i} + w_3 P_{ji} + w_4 C_{ji},$$

where E denotes the sum of four potentials with respect to the different preferences for recommendation. The parameters $w = [w_1, w_2, w_3, w_4]$, $\sum_i w_i = 1$, weights the relative importance of each potential, which can be assigned by the user or fixed arbitrarily by the system. Particularly, if we assign w_1 as 1, it means that we entirely emphasize the result of classification and desire to find out the exact city where the photo was taken.

Then we propose a general framework of tour route recommendation based on the potential model introduced above, as shown in Fig. 5.

Suppose we are interested in the tour of the silkroad. First, users provide a photo they are interested in to ask for tour recommendation. We do not confine the photo to be a Flickr photo associated with tag, since we can predict the tag with its nearest neighborhoods in the visual space. Also, the photo can actually be taken in cities unrelated to the silkroad. It is allowed in the paper since we can recommend the hot city along the silkroad with the most similar sights with the query photo as an alternative. Furthermore, we are going to scale up the number of cities and generalize our model to the recommendation for top cities all over the world. Then the target city will be more likely to be found in our system.

Second, we classify the photo to get the scores that indicate how likely the photo belongs to each city. The

normalized score is what we denote as similarity S_i in (10). It contributes to the final recommendation model in a percentage of w_1 .

Third, we calculate the recommendation potentials of the first query by combining similarity and pre-computed hotness. Note that the proximity and co-visitedness that model the correlation between cities are not considered in the first query, which means that, $w_3 = w_4 = 0$ in (10).

After the first query, we return a sequence of cities in the order of potentials, from high to low. For each city, we display photos of high interestingness in four categories: food, landscape, man-made and person, to show the different characteristics across cities. Then the user chooses a photo from the city that he/she prefers to make the next city recommendation. After the next query photo is chosen, we come back to the second step and update the scores. Unlike in the first query, now we add the proximity and co-visitedness ones into the final recommendation potential, as well as the similarity and hotness. Besides, the city that has been chosen in the former query is no longer considered, since we would not like to take a second visit to the same city. Using the greedy algorithm, we recommend cities by the final potential of each query step by step. Since we have considered the correlation between cities neighboring in sequence by proximity and co-visitedness potentials, the system can finally hand in a reasonable recommendation of tour route for users. A sample of the implementation of this framework is presented in the part D of Section V.

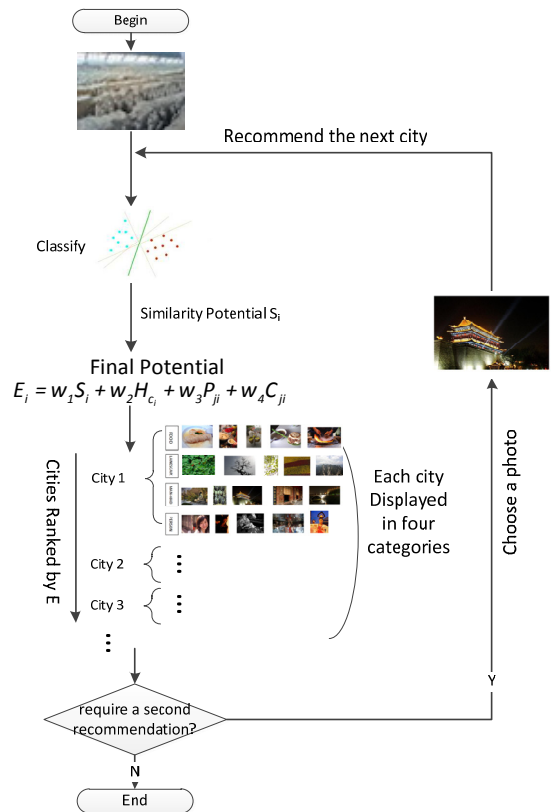


Figure 5. The pipeline of our recommendation framework. The cities returned in each cycle of query are recommended in order, which form the final tour route together.

V. EXPERIMENTAL RESULTS

A. Dataset

Similar to [10], we submit the key word “silkroad” to Wikipedia [9] to find out the names of cities along the silkroad. Then we select 35 key cities and download photos from Flickr with these 35 names using code of [1]. We retain the photo collections of the top 15 cities with 36262 photos in total. The cities and their corresponding sizes of photo collections are listed in Table 1. We can see that the collection of each city is quite unbalanced with the largest size (Cairo) to be 14243 and smallest (Aleppo) to be 136.

Table 1. 15 Cities and their size of photo collection.

City	Size of Photo Collection
Aleppo	136
Ankara	2042
Bagdad	2241
Cairo	14243
Damascus	2507
Dunhuang	513
Gaza	192
Istanbul	4969
Kashgar	573
Kucha	364
Palmyra	349
Samarkand	256
Teheran	2043
Xian	5196
Yazd	638

B. Potentials mined from photo information

1) Hotness of each city

The city hotness computed by (3) that indicates the relative attractiveness of each city is shown in Fig. 6.

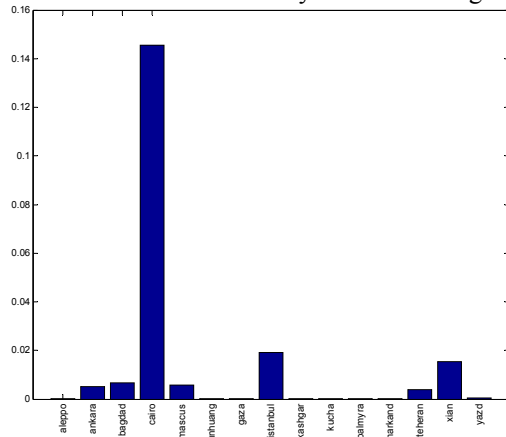


Figure 6. The hotness of each city.

2) Proximity of each two-city pair

It is the normalized reciprocal of distance between two-city pair. Among all the 175 city pairs, the top 5 pairs with largest proximity are listed in Table 2.

Table 2. The top 5 proximity among 175 city pairs.

City pair	Proximity
(Aleppo, Palmyra)	1
(Palmyra, Damascus)	0.9696
(Damascus, Gaza)	0.7444
(Aleppo, Damascus)	0.6751
(Cairo, Gaza)	0.6203

3) Co-visitedness of each two-city pair

Before estimating the co-visitedness potential, we first figure out the hotness of authors. Among the total 3613 authors, the top 10 authors with highest hotness are listed in Table 3.

Table 3. The top 10 hottest authors and their hotness.

Author	Hotness
labanex	1
djtansej	0.4090
ArneSchoell	0.3939
Hubbers	0.3147
vanLyden	0.2976
HikingMatt	0.2120
Patrick Pasenberg	0.1314
monzy	0.1129
buskam7	0.0995
DJ Eddie J	0.0992

For each two-city pair, co-visitedness models the potential for visiting the other city when one of them has been visited. Table 4 shows the 5 city pairs with top co-visitedness.

Table 4. The top 5 co-visitedness among 175 city pairs.

City pair	Co-visitedness
(Bagdad, Teheran)	1
(Teheran, Yazd)	0.8159
(Damascus, Palmyra)	0.7018
(Cairo, Damascus)	0.5861
(Aleppo, Palmyra)	0.5363

C. Classification based similarity potential

The similarity potential is the normalized scores of classification. To perform classification, we equally divide the whole 36262 photos into two subsets, i.e., training set and test set. Since the size of photos from each city are highly unbalanced as shown in Table 1, we first find out the city with the fewest photos and then randomly select the same numbers of photos from the rest of the cities to build the balanced training set. In our experiments, the number of training photos from each city is 69, and thus we have 1035 training photos in total. We implement our learning procedure with RBF kernel SVM by LIBSVM toolbox [11]. We test the effectiveness of the learned classifier in a large scale of 18117 photos, so that we can handle the highly uncertain query photos.

We first employ different coding and pooling strategies to build visual features for classification. As introduced in Section III, we consider both vector quantization and sparse coding as coding strategies, and max pooling and average pooling as pooling strategies. For efficiency, we apply the code of LLC [16] for sparse coding. The classification results are shown in Table 5, from which we find that sparse coding with max pooling achieved the highest classification accuracy. Thus, in the rest of our experiments, we first carry out sparse coding for the SIFT descriptors and then use max pooling to obtain the final visual features.

Table 5. The classification accuracy for various combinations of coding and pooling types.

Accuracy (%)	Vector Quantization	Sparse Coding
Average Pooling	23.42	22.01
Max Pooling	18.60	26.96

In order to evaluate the effects of tags and find the best classification method, we compare the results of four methods, which employ SIFT, TAG, SIFT+TAG, MKL (SIFT+TAG) as their features respectively. Here, we use the training set as reference set. The result is shown in Table 6.

Table 6. The Accuracy of four methods we apply for classification.

Method	SIFT	TAG	SIFT+TAG	MKL (SIFT+TAG)
Accuracy (%)	26.96	19.06	18.95	28.78

The MKL method achieves the best result, which verifies the effectiveness of feature combination. It is worth noticing that the dataset is highly unbalanced and we use only 1035 photos for training while testing on as many as 18117 photos. Since the problem is really challenging, it is no surprise that the classification accuracy is not as good as we usually achieved in other classification tasks. Along with the small training set, the unsatisfying accuracy also results from the inaccurate prediction of visual neighbors. Since the reference set is the training set here and only occupies half of the dataset, it may not be enough to find the real similar photos or cover all the possible tags. To verify this conjecture, we increase the percentage that the reference set occupied in the whole dataset gradually and obtain the corresponding classification accuracies of SIFT, TAG and SIFT+TAG, as shown in Fig. 7. We find that the classification accuracies of both TAG and SIFT+TAG increase with the number of reference photos and are always higher than SIFT only, which indicates that the predication of tags is more accurate when we enlarge the reference set. Thus, we can expect the predicted tags to approach the actual tags if the reference set is large enough, and then the classification accuracy would be much higher as shown in Fig. 7. Besides the prediction of tags, similar photos from different cities also make the classification problem less discriminative. Actually, the photos are difficult to label even by humans. That is an important reason why we should make recommendations more than purely resorting to classification.

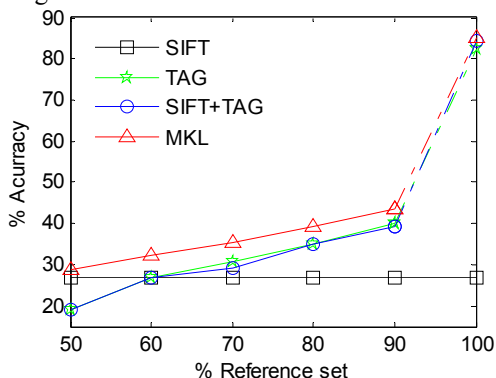


Figure 7. The classification accuracies vary with the percentage that the reference set occupied in the whole dataset. The 100 percentage of the reference set corresponds to the accuracy we obtained when using actual tags.

Furthermore, to show the characteristic of each city, we display photos in four categories when a city is returned. In order to avoid confusion, we address this classification task by choosing from the four methods we

have discussed in city classification. The classification accuracy for four categories is shown in Table 7.

Since this problem is easier, the classification accuracy is much higher than that in Table 6 and the MKL method still outperforms others.

Table 7. The classification accuracy for four categories.

Method	SIFT	TAG	SIFT+TAG	MKL (SIFT+TAG)
Accuracy (%)	65.38	35.84	58.59	65.85

Based on the above comparisons, we continue our subsequent classification work with MKL, which automatically combines visual and textual features.

D. Final Recommendation Model

To test the feasibility and effectiveness of our recommendation model, we begin a sample query with the following photo from **Aleppo**.

Here we do not make too much emphasis on city hotness, and assign it with a weight of 0.1, while the similarity with a weight of 0.9. (Of course, these weights can be set by user to satisfy his preference). Then we gain the final potentials shown in Table 8.



Figure 8. The first query photo.

Table 8. The final potentials for the first query.

City	Final Potential
Aleppo	0.9000
Cairo	0.3092
Bagdad	0.2719
Ankara	0.2622
Kashgar	0.2589
Palmyra	0.2589
Samarkand	0.2587
Teheran	0.2582
Istanbul	0.2117
Kucha	0.1824
Damascus	0.158
Gaza	0.1105
Dunhuang	0.0999
Xian	0.072

It is inspiring that we have recommended the right city Aleppo. So we add Aleppo as the first city into the tour route. The next step is to make the second query with a new photo from Aleppo as shown in Fig. 9.

To evaluate the influence of different weights, we set weight vector w as $[0.1, 0.1, 0.6, 0.2]$, $[0.1, 0.2, 0.1, 0.6]$ and $[0.6, 0.2, 0.1, 0.1]$ respectively. The weight vector consists of similarity, hotness, proximity and co-visitedness from left to right. The final potentials varying with three weights are shown in Table 9.

Note that we do not consider the potential of Aleppo any more since it is already in the tour route. We first emphasize the potentials of proximity and co-visitedness with $w=[0.1, 0.1, 0.6, 0.2]$ and then Palmyra is returned as the recommendation. The reason is easy to find from Table 2 and Table 4 that Palmyra is the nearest city to Aleppo and they are most commonly co-visited. When

Table 9. The final potentials vary as we change the weights.

$w=[0.1, 0.1, 0.6, 0.2]$		$w=[0.1, 0.2, 0.1, 0.6]$		$w=[0.6, 0.2, 0.1, 0.1]$	
City	Final Potential	City	Final Potential	City	Final Potential
Palmyra	0.7877	Cairo	0.5392	Bagdad	0.6391
Damascus	0.5106	Palmyra	0.5024	Palmyra	0.6361
Cairo	0.3814	Damascus	0.3802	Cairo	0.5596
Bagdad	0.2782	Bagdad	0.148	Teheran	0.5018
Ankara	0.2666	Istanbul	0.124	Istanbul	0.4959
Gaza	0.2548	Samarkand	0.1138	Samarkand	0.4939
Istanbul	0.2284	Teheran	0.1016	Dunhuang	0.4652
Teheran	0.1804	Yazd	0.087	Yazd	0.4563
Yazd	0.1499	Ankara	0.0819	Ankara	0.2679
Samarkand	0.1365	Dunhuang	0.0812	Gaza	0.2627
Dunhuang	0.1022	Gaza	0.0739	Xian	0.1436
Kashgar	0.0632	Kashgar	0.0494	Damascus	0.1263
Xian	0.0501	Xian	0.0443	Kashgar	0.118
Kucha	0.0493	Kucha	0.023	Kucha	0.1111

the weight is changed to $[0.1, 0.2, 0.1, 0.6]$, which means we highly emphasize city hotness and concern about co-visitedness, then Cairo, a definitely hot city, is retrieved with the highest potential. If we emphasize the visual similarity and hotness with $w=[0.6, 0.2, 0.1, 0.1]$, we get Bagdad as the second recommendation. In Bagdad we can find buildings resemble the query photo displayed in Fig. 9, such as the two photos in Fig. 10. We can conclude from the above discussions that the potential weights reflect user's preference and in order to meet this preference, the recommendation results vary with the change of weights. However, we can also observe that although the final potential changes, the cities with top potentials are relatively constant, i.e., Palmyra, Cairo and Bagdad are always among the top 4 cities.

Since we will rank the cities according to their potentials and photos of each city are displayed in four categories ranked by interestingness, users can easily find the photo that interests them to continue generating the next city in the tour route.

In this example, we choose the right-hand side photo of Fig. 10 as the third query photo.

To generate a tour route, we would like the next city to be near Bagdad, so we emphasize the proximity potential by a weight 0.5 and set the weight of similarity, hotness, co-visitedness to 0.1, 0.2, and 0.2 respectively. Then we get the final potentials listed in Table 10.



Figure 9. The second query photo.



Figure 10. Photos from Bagdad that resemble the second query photo from Aleppo.

Table 10. The final potentials for the third query.

City	Final Potential
Teheran	0.4441
Cairo	0.331
Palmyra	0.3245
Ankara	0.2888
Yazd	0.2855
Aleppo	0.2193
Istanbul	0.217
Gaza	0.2018
Damascus	0.1938
Samarkand	0.1684
Xian	0.1566
Kashgar	0.1104
Kucha	0.0703
Dunhuang	0.0444

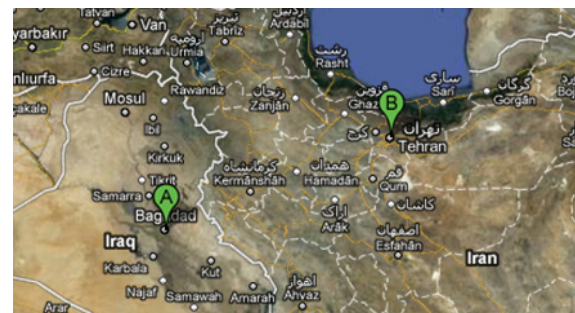


Figure 11. Location of Teheran and Bagdad in Google Map with Bagdad denoted as A, and Teheran denoted as B.

Note that we do not evaluate the potential of Bagdad, since we do not want to visit the same city for a second time. In this query, we get Teheran as the first preference, which is located in Iran, next to Iraq where Bagdad is located. And the distance between them is only 693.74-km, which takes less than one hour to fly from the one to another. Their relative locations in Google Map [14] are displayed in Fig. 11.

Also, in the collection of Teheran, we find photos resemble the third query photo from Bagdad, such as the two photos shown in Fig. 12.

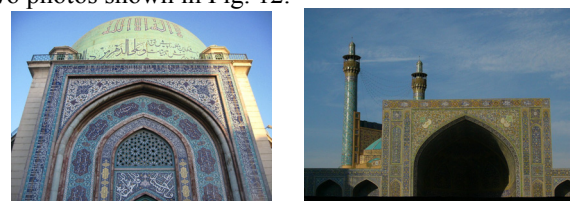


Figure 12. Photos from Teheran which resemble the third query photo.

What's more, Teheran and Bagdad is the most commonly co-visited city pair as shown in Table 4. We suppose that authors tend to visit both of them for the reason that they are near in terms of location and are both attractive. As a result, the authors who visited the two cities during a relatively short time have actually acted as our role-model authors who help us make this recommendation for users.

According to the above analyses, we suppose that Teheran is the best choice for the third query with weights [0.1, 0.2, 0.5, 0.2], which emphasize on proximity potential and less concern with similarity.

Now if we accept the above recommendation results, we can obtain our three-city route as "Aleppo→Bagdad→Teheran", and we can continue by choosing a photo from Teheran to ask for the fourth city. The cycle of query can be terminated any time when we get a tour route as long as we expect. Finally, a complete tour route is generated as "Aleppo→Bagdad→Teheran→...→...". Interestingly, if we continue by placing emphasis on proximity potential alone, we will get one of the possible tour routes the same as a fragment of the original silkroad, as illustrated in Fig. 13.

VI. CONCLUSION

In this paper, we propose a general framework of making recommendations for tourists who are interested in photos and have no idea where they were taken. We believe that the user is not necessarily interested in the exact city where the photo was taken. Instead, what the user is more interested is the style of the scene in the photo. Therefore, unlike traditional methods, we perform classification only to obtain the resulting scores as one of our four potentials, i.e., similarity potential. Besides, we model other three potentials, including hotness, proximity and co-visitedness, by analyzing the information attached to each photo. Making an overall consideration of all the four factors, we can make a final recommendation of a tour route for users step by step. In each step, we display photos in four categories, which show the characteristics of sights and culture of each city. The sample query shows the effectiveness and reasonableness of our model for tour route recommendation. One of our future works is to explore a more principle way to model the interactive information among photos and to combine them more sophisticatedly. Besides, we would like to

increase the accuracy of classification and generalize our framework into other settings such as recommendation for top-100 cities all over the world.

ACKNOWLEDGMENT

The work of this paper is partially supported by NSFC funds (60873085 and 61103060). We would like to thank Xiaozhen Qi for helpful discussions, and Yuefeng Chen, Honghui Wang et al., who helped us with photo labeling.

REFERENCES

- [1] J. Hays, A. A. Efros. "IM2GPS: estimating geographic information from a single image," Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2008.
- [2] Y. Li, D.J. Crandall, D.P. Huttenlocher. "Landmark classification in large-scale image collections," Proceedings of IEEE 12th Int. Conf. on Computer Vision, pp.1957-1964, Sept. 29 2009-Oct. 2, 2009.
- [3] M. Guillaumin, J. Verbeek, C. Schmid. "Multimodal Semi-supervised Learning for Image Classification," Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2010.
- [4] G. Wang, D. Hoiem, D. Forsyth. "Building text features for object image classification," Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2009.
- [5] <http://www.flickr.com>
- [6] A. N. Langville, C. D. Meyer. "Survey: Deeper inside PageRank," Internet Mathematics, 1(3): 335-380, 2003.
- [7] D.G. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints," IJCV, 60(2):91-110, 2004.
- [8] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet. "SimpleMKL," JMLR, 9:2491-2521, 2008.
- [9] <http://www.wikipedia.org>
- [10] Q. Wang, X. Qi, and J. Xu. "e-Silkroad: A Sample of Combining Social Media with Cultural Tourism," Workshop on connected multimedia, Proceedings of ACM Int. Conf. on Multimedia, 2010.
- [11] C.C. Chang, C.J. Lin. "LIBSVM: a library for support vector machines," <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, 80:604-611, 2001.
- [12] Y-L. Boureau, F. Bach, Y. LeCun, J. Ponce. "Learning Mid-Level Features for Recognition," Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), San Francisco, USA, pp. 2559-2566, 2010.
- [13] Y-L. Boureau, J. Ponce, Y. LeCun. "A Theoretical Analysis of Feature Pooling in Visual Recognition," Proceedings of 27th Int. Conf. on Machine Learning



Figure 13. The illustration of the sample tour route. The solid line which connects Aleppo, Bagdad and Teheran denotes the tour route we obtain currently and the dash line forms a possible tour route which extends exactly as a fragment of the original silkroad.

(ICML), Haifa, Israel, 2010.

- [14] <http://maps.google.com>
- [15] J. Yang, K. Yu, Y. Gong, T. Huang. "Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification," Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2009.
- [16] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong. "Locality-Constrained Linear Coding for Image Classification," Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), San Francisco, USA, pp.3360-3367, 2010.
- [17] Y. Yang, D. Xu, F. Nie, J. Luo, Y. Zhuang. "Ranking with local regression and global alignment for cross media retrieval," Proceedings of ACM Int. Conf. on Multimedia, pp. 175-184, 2009.
- [18] Y. Yang, Y. Zhuang, F. Wu, Y. Pan. "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval," IEEE Trans. on Multimedia, 10(3):437-446, 2008.
- [19] Y. Yang, Y. Zhuang, W. Wang. "Heterogeneous Multimedia Data Semantics Mining using Content and Location Context," Proceedings of ACM Int. Conf. on Multimedia, 2008.



Xiujun Chen received her B.S. degree in computer science from Northwestern Polytechnical University in 2011. She is currently a postgraduate in the School of Computer Science and Engineering at Northwestern Polytechnical University, China. Her research interests include computer vision, image classification, machine learning, etc.



Qing Wang (correspondence author) received his B.S. degree in Information Mathematics from Department of Mathematics, Peking University in 1991, M.S. degree in Pattern Recognition and Intelligent Control in 1998, and PhD degree in Computer Science and Technology in 2000 from Department of Computer Science and Engineering,

Northwestern Polytechnical University, China, respectively. He is currently a professor in the School of Computer Science and Engineering, Northwestern Polytechnical University, China.

He worked as research assistant from Oct. 1999 to April 2000 and research fellow from Nov. 2000 to Feb. 2002 at Center for Multimedia Signal Processing, Department of Electronics and Information Engineering, Hong Kong Polytechnic University. He worked as visiting research fellow from Nov. 2003 to May 2004 in the School of Information Technology, University of Sydney, Australia, and as visiting professor from March 2009 to Sep. 2009 in School of Computer Science, Carnegie Mellon University, USA, respectively. He has published more than 100 technical papers. He acted as paper reviewer for international journals, such as Pattern Recognition, Pattern Recognition Letter, IEEE Tran. on Signal Processing, IEEE Tran. on Image Processing, and so on. He is a senior member of China Computer Federation (CCF) and a member of IEEE and ACM. His research interests include 3D scene reconstruction, image based modeling and rendering, light field theory and application, computational photography, etc.

Guest Editors' Introduction: Special Section on ISIP 2011

Fei Yu

Peoples' Friendship University of Russia, Russia
Email: hunanyufei@126.com

Yiqin Lu

South China University of Technology, China
Email: eeyqlu@scut.edu.cn

Chin-Chen Chang

National Chung Hsing University, Taiwan
Email: ccc@cs.ccu.edu.tw

Yan Gao

Henan Polytechnic University, China
Email: gaoyan@hpu.edu.cn

This special section comprises of nine selected papers from the fourth International Symposium on Intelligent Information Technology and Security Informatics (ISIP 2011). The conference received 310 paper submissions from 11 countries and regions, of which 157 papers were selected for presentation after a rigorous review process. From these 157 research papers, through two rounds of reviewing, the guest editors selected nine as the best papers on the Multimedia track of the Conference. The candidates of the Special Issue are all the authors, whose papers have been accepted and presented at the ISIP 2011, with the contents not been published elsewhere before.

The ISIP 2011 was co-sponsored by Henan Polytechnic University, China; Peoples' Friendship University of Russia, Russia; Feng Chia University, Taiwan; Jiangxi University of Science and Technology, China; Fudan University, China; South China University of Technology, China; Nanchang Hang Kong University, China; Jiaying University, China; Academy Publisher of Finland, Finland and took place on August 19-22, 2011, in Jiaozuo, China.

“Speech Separation in the Vehicle Environment Based on FastICA Algorithm”, by Jindong Zhang, Guihe Qin and Ye Liu. In this paper, the FastICA algorithm in signal process and statistics was studied and used to separate the driver's speech in the vehicle environment, and realizes the pretreatment in recognizing driver's speech.

“Semantic Analysis of Traffic Video Using Image Understanding”, by Jian Wu, Zhi-ming Cui, Heng-jun Yue, and Guang-ming Zhang. This paper introduces the methods of image understanding and proposes a video semantic analysis framework using scene analysis for recognizing some semantic events in the traffic video.

“Research on Video Quality Assessment”, by Chunting Yang, Yang Liu and Jing Yu. In this paper, research has been focused on developing novel objective evaluation metrics which enable prediction of the perceived quality level. The temporal correlations of video frames and the visual interest feature are considered in this method. Meanwhile the metrics are capable of capturing spatial distortions in video sequences.

“Tele-Immersive Interaction with Intelligent Virtual Agents Based on Real-Time 3D Modeling”, by Shujun Zhang and Wan Ching Ho. In this paper author explored the combination of real-time 3D modelling and HAI and proposed a prototype of mixed reality system for tele-immersion. Image-based real-time 3D modeling techniques offer essential ways to support tele-immersion and natural HAI.

“The Research of Image Encryption Algorithm Based on Chaos Cellular Automata”, by Shuiping Zhang and Huijune Luo, In this paper, the Research puts forwards an image encryption algorithm based on chaos cellular automata, to take full use of the two-dimensional chaotic system's extreme sensitivity to initial conditions and the features with parallelism, complexity and randomness of two-dimensional reversible cellular automata, to make the encrypted key, encrypted image and original image much more have complexity, randomness and unpredictability.

“Improved MFCC Feature Extraction Combining Symmetric ICA Algorithm for Robust Speech Recognition”, by Huan Zhao, Kai Zhao, He Liu and Fei Yu. This paper proposed using symmetric orthogonalization in ICA for projecting log Mel spectrum into a new feature space as a substitute in extracting speech features to solve the problem of cumulative error and unequal weights that deflation orthogonalization brings, so as to improve the robustness of speech recognition systems, and increase the efficiency of estimation at the same time.

“A Watermarking Technique based on the Frequency Domain”, by Huang-Chi Chen, Yu-Wen Chang and Rey-Chue Hwang. In this paper, a modified algorithm is presented to improve the defect of the JPEG quantification in order to reduce the bit error rate (BER) of the retrieved watermark.

“Image Copy-Move Forgery Detecting Based on Local Invariant Feature”, Li Jing, and Chao Shao. This paper firstly analyzes and summarizes block matching technique, then introduces a copy-move forgery detecting method based on local invariant feature matching.

“OWL-S based Service Composition of Three-dimensional Geometry Modeling”, by Jiangning Yu, Hongming Cai and Ailing Liu. This paper proposes an OWL-S framework for distributed CAD system based on the combination of semantic web service and CAD technology.

We would like to take this opportunity to thank the authors for the efforts they put in the preparation of the manuscripts and for their valuable contributions. We wish to express our deepest gratitude to the program committee members for their help in selecting papers for this issue and especially the referees of the extended versions of the selected papers for their thorough reviews under a tight time schedule. Last, but not least, our thanks go to the Editorial Board of the Journal of Multimedia for the exceptional effort they did throughout this process.

In closing, we sincerely hope that you will enjoy reading this special issue.



Fei Yu was born in Ningxiang, China, on February 06, 1973. Before Studying in Peoples' Friendship University of Russia, Russia, He joined and worked in Hunan University, Zhejiang University, Hunan Agricultural University, China. He has wide research interests, mainly information technology. In these areas he has published above 90 papers in journals or conference proceedings and a book has published by Science Press, China (Fei Yu, Miaoliang Zhu, Cheng Xu, et al. Computer Network Security, 2003). Above 70 papers are indexed by SCI, EI. He has won various awards in the past. He served as many workshop chair, advisory committee or program committee member of various international ACM/IEEE conferences, and chaired a number of international conferences such as IITA'07, IITA'08; ISIP'08, ISIP'09, ISIP'10, ISIP'11; ISECS'08, ISECS'09, ISECS'10, ISECS'11; WCSE'08, WCSE'09, WCSE'10, WCSE'11 and ISISE'08, ISISE'09, ISISE'10. He have taken as a guest researcher in State Key Laboratory of Information Security, Graduate School of Chinese Academy of Sciences, Guangdong Province Key Lab of Electronic Commerce Market

Application Technology, Jiangsu Provincial Key Lab of Image Processing and Jiangsu Provincial Key Laboratory of Computer Information Processing Technology.



Yiqin Lu was born in 1968, Deqing, China. He obtained his Ph.D degree in circuit and system from South China University of Technology in 1996.

He has been a post-doctor in Dept. of Computer Science in City University of Hong Kong during 1997-1998. Now he is a professor of South China University of Technology. He has supervised 17 research projects, and published one book and more than 60 journal or international conference papers. His research interests include telecommunications networks, computer networks, home networks, and the theories and application of Petri nets.

Prof. Lu is now the director of the South China Network Center of China Education and Research Network, the director of Information and Network Engineering and Research Center of South China University of Technology. He is an IEEE member. He owned the first grade prize as the Excellent Research Student of Guangdong Province in 199, and the third grade Science and Technology Award of Guangdong Province in 2004.

Student of Guangdong Province in 199, and the third grade Science and Technology Award of Guangdong Province in 2004.



Chin-Chen Chang was born in Taichung, Taiwan on Nov. 12th, 1954. He obtained his Ph.D. degree in computer engineering from National Chiao Tung University. He's first degree is Bachelor of Science in Applied Mathematics and master degree is Master of Science in computer and decision sciences. Both were awarded in National Tsing Hua University. Dr. Chang served in National Chung Cheng University from 1989 to 2005. His current title is Chair Professor in Department of Information Engineering and Computer Science, Feng Chia University, from Feb. 2005.

Prior to joining Feng Chia University, Professor Chang was an associate professor in Chiao Tung University, professor in National Chung Hsing University, chair professor in National Chung Cheng University. He had also been Visiting Researcher and Visiting Scientist to Tokyo University and Kyoto University, Japan. During his service in Chung Cheng, Professor Chang served as Chairman of the Institute

of Computer Science and Information Engineering, Dean of College of Engineering, Provost and then Acting President of Chung Cheng University and Director of Advisory Office in Ministry of Education, Taiwan.

Professor Chang has won many research awards and honorary positions by and in prestigious organizations both nationally and internationally. He is currently a Fellow of IEEE and a Fellow of IEE, UK.



Yan Gao, male, born in 196, Ph.D., professor, the senior member of Senior Member of China Computer Federation and China Education Ministry Steering Committee Member of Computer Education on Arts. The director of Educational Committee of Henan provincial Computer Federation, the deputy dean of computer and software academe of Henan Polytechnic University.

He current research interests include Intelligent Control, Intelligent Information Processing. He has directed (or participated) and completed more than 20 items of the research projects. More than 40 research papers have been published in domestic and foreign academic journals or conferences. 17 of them have been indexed by SCI/EI/ISTP. As the directed or associate editor, 3 treatises and 6 teaching materials have been published.

Speech Separation in the Vehicle Environment Based on FastICA Algorithm

Jindong Zhang

College of computer science and technology, Jilin University, Changchun, China
Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, Changchun, China
Email: zhangjindong_100@163.com

Guihe Qin and Ye Liu

College of computer science and technology, Jilin University, Changchun, China
Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, Changchun, China
Email: qingh@jlu.edu.cn

Abstract—The speech interaction in-vehicle was mainly realized by the speech recognition. The human-machine interaction around was usually disturbed by the noise, and the speech received by the receiver was not the original pure speech, so compared to the pure environment, the accuracy of the speech recognition declined so sharply that it could not meet the demand of the practical application of human-machine interaction. So the speech recognition was required to have the strong adaptability and processing capacity to the speech with noise. In this paper, the FastICA algorithm in signal process and statistics was studied and used to separate the driver's speech in the vehicle environment, and realizes the pretreatment in recognizing driver's speech. The effectiveness of the method had been validated in actual vehicle experimental configuration.

Index Terms—integrated voice-data communication, intelligent control, speech processing, human-machine interfaces, fast independent component analysis (Fast ICA)

I. INTRODUCTION

With the development of science and technology and the higher requirement of vehicle performance, vehicles had more and more devices, and the human-machine interaction became more and more important. Considering the driving safety, the drivers had to focus on the road status and keep their hands on the steering wheel. If drivers could control the GPS navigation system through the speech, such as setting destinations, searching places, and controlling et al., they could avoid using their hands and eyes, and this had very important meanings for freeing the drivers' hands. Using speech as a means of human-machine interaction could avoid their dependence on hands, and it also had very important practical meanings on convenience and safety.

The most significant advantage of the speech communication interface was to keep the drivers' hands on the steering wheel and eyes on the road. To realize the speech recognition technology, it needed very high performance processor and software. The condition was that the hardware already existed, but the software needed to be improved. Many companies such as Lernout & Hauspie (L&H), Nuance, IBM, Dragon and Speech

Works had already developed speech recognition software which could run on various platforms [1]. The key technologies of the speech recognition system were creating the command vocabulary and training the vocabulary. And it required high recognition accuracy. When the vocabulary was large or there were more users the accuracy would decline. In fact in vehicle, the vocabulary was small and the user was only the driver, so the emphasis of the speech recognition system was to improve the performance of the system through improving the key technology.

Delphi Automotive Systems had provided hand-free speech recognition system since 1996, and it also provided telematics series products cooperated with BMW. The navigation service such as Jaguar ASSIST, Wingcast and Wireless-Car provided by Ford Company was used in Explorer, Mercury Sable and Jaguar Sedan [2]. The Sarah Auto PC, which showed in Paris Motor Show by the Citroen in 1998, starts a new situation in vehicle communication technology [3]. It based on Microsoft WINCE2.0 operating system, controlled by speech, reacted to driver's voice command and dialogize with handlers through speech synthesis system. In the same year, the General Motors Company vigorously developed its Onstar electronic system, and made it a speech recognition system from the basic lane driving assistant system. In 2000, the General Motors Company produced the first speech controlling wireless internet automobile. Simultaneously, some limousine such as "Jaguar S Series" installed the simple speech control system, through which the driver could regulate the air-conditioning and sound by speech instruction. In 2001, the Audi AG showed the Multi Media Interface(MMI) control technology in Frankfurt Motor Show, which integrated the electronic system in the car and realized the speech control. In 2003, Dr. Sapphire Company developed the vehicle audio multimedia, which controlled telephone, sound and navigation system by speech recognition. In the same year, the Mercedes-Benz's new generation vehicle type of E-Class used speech recognition control technology to control the navigation system and the vehicle phone. In 2004, the scientists in

Waston study center of IBM Company realized the vehicle speech control system by using the video channel adjustment to supplement the speech input. In 2005, technicians in Toyota Company developed the speech control system for Crown Royal Saloon G. In the same year, the new U style concept vehicle of Ford showed the most advanced announced speech control technology so far, which allowed people to manipulate the vehicle systems included entertainment, navigation, mobile phone and air-conditioning control by natural speech [4]. The speech recognition laboratory in Electronics Department of Tsinghua University in China had done a lot of works in this aspect. They designed an independent speech recognition model, which included two kernel parts and simple speech input and output devices. The kernel parts were composed of UniSpeech produced by Infineon Company and a flash to store data and program [5].

The speech interaction in-vehicle was mainly realized by speech recognition. The speech recognition was a technology to implement corresponding control through recognizing speaker's speech feature or identification to correctly judge the speech connotation. It was the key technology of using speech in human-machine interaction. In recent years, it was used in many fields. Depending on speech recognition hardware and tool software to recognize speaker's voice, it controlled corresponding actuators. At present, the speech recognition in laboratory quiet environment was very mature, and the recognition accuracy could reach 90%, which could satisfy the normal application. These systems were large vocabulary continuous speech recognition system such as Via Voice of IBM, Whisper of Microsoft and SPHINX.H of CMU [6]. But the coming multimedia age urgently required the speech recognition system to use practically, and this required the speech recognition to have the strong adaptability and processing capacity to the speech with noise.

Because the human-machine speech interaction in-vehicle disturb by the noise in environment, the received speech was already not the original pure speech. Compared to the pure environment, the recognition accuracy decreased sharply, and could not meet the demand of practical human-machine interaction. The difficulty of the speech recognition was how to solve the noise. Now there was much study about noise eliminating. The common methods could be summarized in four aspects: spectral subtraction, environment structured technology, correcting recognizer model (not correcting speech signal) to adapt noise, and establishing noise model. But these methods could not eliminate the noise influence completely, and they needed further improvement. This kind of speech recognition system in the paper distinguished the specific speaker, and was often very accurate. It demanded user to train the speech recognition system to get the recognition result. Realizing noise eliminating and suppression, enhancing the speech were the basic and key contents for improving speech recognition accuracy. After the front end processing on the speech inside vehicle with noise, such as noise

eliminating and speech enhancing, the speech as pure as possible was got, and then the speech recognition in noise environment was done by using the mature speech recognition technology [7].

In 1990, Bregman proposed the computer separation method which aimed at separating speech signal in his published book [8]. In 1995, after the founding of the first Computational Auditory Scene Analysis (CASA) study team, the computer speech separation aiming at speech became a hot research. In the same year, A. J. Bell and T. J. Sejnowski published the milestone literature in ICA theory and technology developing. Then scholars began to use ICA technology to separate the synthetic speech signal with noise, and ICA technology began to use in multi-channel speech signal separation [9]. But ICA separation algorithm existed limitations in actual environment. In October, 2003 at Montreal in Canada, a speech separation prospect study and discussion team initiated and composed by National Natural Science Foundation, and other fund committee also participated. The team communicated with all speech study fields and focused on discussing the future of the speech separation. In November 2004, at Quebec in Montreal, a study and discussion team about speech separation and understanding in complex speech environment initiated by Air Armed Forces Natural Science Research Association and National Natural Science Foundation [10]. But ICA algorithms still have problems in the following aspects.

a) the estimation of the number of blind source signal. At present the speech blind separation algorithm assumes that the number of the observed mixture signals is equal to or greater than the number of the source signals. For unknown source signals and the number of the observed signals less than the source signals, it is difficult to analyze and needs further study.

b) The space location of the source signals. When the sensors' position of the source signals and observed signals are not in the same plane but have the spatial relationship, there will be energy masking and information masking, and then the space location only can be done by the limited information in the observed signals.

Therefore, studying and proposing the front end processing technology in the driving environment was very important. Based on the FastICA algorithm in the signal processing and statistics, the paper used the method to separate the driver's speech, and realized the front end processing to recognize the driver's speech.

The remaining parts of this paper were organized as follows: Section II described the FastICA algorithm. Section III described the configuration of experiments in the actual vehicle environment, analyzed the experimental results, and the conclusions were given in Section IV.

II. ICA AND FASTICA ALGORITHM

As we know that independent component analysis (ICA) was a popular technology for solving the BSS problems. Paper [11] addressed some of the first works to

apply ICA in symbol demodulation. ICA relied on higher-order statistics which were typically the fourth-order statistic kurtosis to solve the BSS problems. Papers [12] and [13] addressed the issue of delay estimation with ICA. Jammer mitigation with a BSS principle was first discussed in [14], where an ICA method called JADE and second-order methods were used to mitigate a temporally correlated jammer. The paper [15] also addressed the issue of blind beam forming to jammer mitigation. These methods were very popular in other domains, and they were attracted attention in the field of communication engineering. Some person thought the reason was that some of these methods had an artificial neural network background.

In the vehicle environment we introduced ICA technology into speech separation. Combining an ICA element to standard techniques enabled a robust and computationally efficient structure. In the paper, we introduced a switching techniques based on BSS/ICA effectively to combat interference [14]. BSS was an idea of signal processing where mixtures of several sources were separated without the knowledge of the mixing processes. Several schemes existed for interference suppression based on a similar principle. Considering the following linear model, the popular frame work for solving the BSS problem was ICA.

$$M_{orig} = AO + v \tag{1}$$

Usually, we must make the fundamental restrictions before the mixed-signal separation. First, the components of O are statistically independent. Second, at most one component of O is Gaussian distributed. Third, the mixing matrix A is full rank. In the model the linear mapping was called the mixing matrix. The model assumed some noise v considered to be Gaussian. Solution to the linear source separation problem was not possible, if there was no information on some of the variables A or O , in addition to the observed data M_{orig} .

$$M_{orig} = \begin{bmatrix} M_{orig,1} \\ M_{orig,2} \\ \vdots \\ M_{orig,K} \end{bmatrix}, O = \begin{bmatrix} O_1 \\ O_2 \\ \vdots \\ O_N \end{bmatrix}, v = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_K \end{bmatrix} \tag{2}$$

$$O_i = [O_i(1) \dots O_i(t)] \tag{3}$$

The symbol t represented time, but could represent some other variable, e.g., space. The model consisted of N sources of t samples, i.e.

$$M_{orig,i} = [M_i(1) \dots M_i(t)] \tag{4}$$

Usually, the observations M_{orig} consisted of K mixtures of the sources. And it was assumed that there were at least as many observations as sources i.e., $K \times N$.

$$A = [a_1 \ a_2 \ \dots \ a_N] \tag{5}$$

If the mixing matrix A was known and the noise was negligible, the sources could be estimated by finding the matrix B , the inverse of the mixing matrix A . The full-rank assumption was the necessary and sufficient condition for the existence of the pseudo-inverse of A .

$$BM_{orig} = BAO = O \tag{6}$$

If there were as many observations as sources, then A was square and has full-rank. So that, $B=A^{-1}$. When there were more observations than sources, the existed several matrices B that satisfy the condition $BA = I$. In this case, the choice of B depended on the components of O which we were interested in. For cases where there were less numbers of observations than sources, a solution did not exist unless further assumptions were made. Now the rank of A was less than the number of sources. There were some redundancies in the mixing matrix, and hence further information was required.

The algorithm could achieve optimization by regulating separation matrix which used the stochastic gradient algorithms. On the assumption that the independent component analysis data model satisfied the restrictions, the convergence rate of FastICA was at least quadric. With the help of nonlinear functions, FastICA finds the independent components of non-Gaussian distribution directly. The characteristic of FastICA ensured that it could achieve optimization by a suitable nonlinear function; especially it could gain the algorithm which was robustness.

We processed BSS for the system using FastICA algorithm in the paper as figure 1, the detailed execution was as follows.

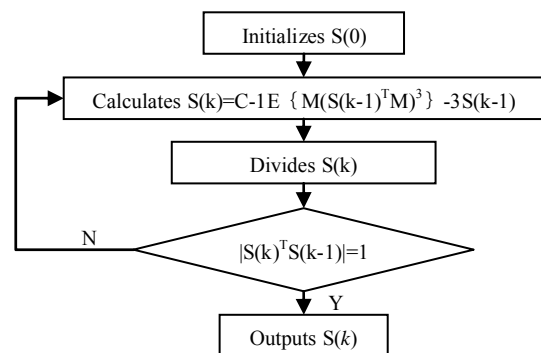


Figure 1. The process of algorithm

Step1 Initializes the separation matrix $S(0)$, sets its mold to 1, $k=1$;

Step2 Calculates $S(k)=C-1E \{ M(S(k-1)^T M)^3 \} - 3S(k-1)$, the expected value may be calculated out by the sample values of M vector;

Step3 Divide $S(k)$ with $\|S(k)\|$;

Step4 If $|S(k)^T S(k-1)|$ don't approach 1 closely, then, makes $k=k+1$, returns to step 2. Otherwise, outputs $S(k)$.

$S(k)$ was calculated by the algorithm equivalents to one of the columns of mixed orthogonal matrix O . This meant that we separated out a nonlinear Gaussian signal

$S(k)^T M(t)$ by $S(k)$, t denotes the time. This signal was one of the speech signals. In order to estimate out N independent components, the algorithm must be run N times. To ensure that each was different estimates of the independent components, an operation of orthogonal project was needed. Because the columns of the matrix O were orthogonal, therefore, the independent components could be estimated one by one. From step 3,

where,

$$S(k) = S(k) - OO^T S(k) \quad (7)$$

We projected the value of current $S(k)$ to the columns of the mixed matrix O , added a projection operation. Then divided $S(k)$ with $\|S(k)\|$. The initial random vectors at the beginning also run the project before the implementation of recursion. In order to avoid deterioration of estimated $S(k)$, the projection could be cancelled after a certain number of iterations. Therefore, we combined with the FastICA algorithm, optimized the separation matrix W by the stochastic gradient algorithm. This algorithm processed the iterations by the determinations of the greatest negative entropy, gained the following formula:

$$S_j(k) = \frac{S_j(k)}{\|S_j(k)\|} E[M_j G(S_j^T(k) X_j)] - E[G'(S_j^T(k) M_j)] S_j(k) \quad (8)$$

Where, $S_j(k)$ was the row vector corresponding to the j th speech signal within matrix S that iterated k times. Continue the iteration like this, until all signals were separated. After each iterations we must carry on processing to $S_j(k)$, to ensure the results which been separated had the unit energy. If two neighboring $S_j(k)$ s were equal to each other or have small difference, then the iterative procession would terminates. Besides, after separating a speech signal each time, we must remove this speech signal from the composite signal.

III. EXPERIMENTAL RESULTS

To handle the noise disturbing to the speech recognition system in vehicle environment, we needed the sound data in vehicle environment to provide the basis for the subsequent speech recognition. According to the vehicle structure and working characteristics, we analyzed the noise sources and made the concrete experiment environment under the new automobile industry national standard <GB-T18697-2002 Acoustics-Measurement of noise inside motor vehicles>.

In this experiment, the distance between the vehicle and large objects was longer than 20m, so the vehicle sound radiated could only become part of the inside vehicle noise by the reflection of the road but not the buildings, walls and other similar large objects. The temperature must between -50°C and $+350^\circ\text{C}$. The wind speed along the measurement road at the height of 1.2m must never exceed 5m/s. And other meteorological conditions mustn't influence the measurement result. The noise in-vehicle was greatly influenced by the roughness of the road surface condition, and the flat road surface

could generate smooth noise inside the vehicle [16]. So in the experiment, the selected road was flat asphalt road without joints and unevenness. The road surface was dry and had no sundries such as snow, dirt, stones and leaves. Then there was no way to increase the sound level inside vehicle. The vehicle was no-load in the experiment while collecting noise, there was no other load except the driver, testers and test devices. In the experiment, the skylight, all the windows, air inlets, air outlets and assist devices such as windshield wiper, heater devices, fans, and air conditioning were closed as they were not the main sources of the noise inside vehicle. The adjustable seat should adjust to the middle of the horizontality and vertical. The back of the seat should be vertical. And the adjustable headrest should be in the middle. The specific experiment environment settings list in Table I.

TABLE I.
EXPERIMENT ENVIRONMENT SETTINGS

Measurement Date: July 16, 2008
Weather: Cloudy Day
Measurement Place: high-tech zone, Changchun, China
Temperature($^\circ\text{C}$): 22 celsius degree
Road Situation: Flat asphalt road
Wind Speed(m/s): Southerly, level 1.0, 0.5m/s
Vehicle Style: BORA1.6, M1
Manufacture Date: 1985-10
Engine Type: 4 Cylinder
Drived mileage(km): 33000
Type: BJH026729
Rated Passengers No or Maximum Total Mass (kg): 5passengers/1830
Rated Power Capability (kw): 74
Rated Speed (r/min): 2500
Transmission: Manual Transmission, 5 gears
Microphone Style: BOM6022HL20-J263-C1033
Sensitivity Range: $-26\pm 3\text{dB}$, $RL= 2.2\text{K}\Omega$, $V_s=3.0\text{V}$ (DC) (1KHz)
$0\text{dB}=1\text{V/Pa}$) Impedance: Max. $2.2\text{K}\Omega$ 1KHz ($RL=2.2\text{K}\Omega$)
Frequency: 50-16000 Hz
Current Consumption: Max $500\mu\text{A}$, $RL=2.2\text{K}\Omega$, $V_s=3.0\text{V}$ (DC)
Operation Voltage Range: 2.2V-5.0V (DC)
Max. Sound Pressure Level: 115dB S.P.L
S/N Ratio: More than 58dB 1kHz, $0\text{dB}=1\text{V/Pa}$, A-weight
Sensitivity Reduction: 3.0V-2.2V, Sensitivity Variation less than 3dB
Sound level Meter Style: TDJ824
Accuracy Level: $\pm 1.5\text{dB}$ 94dB@1kHz
Verification Effective Date: July 1, 2008
Two Laptops: IBM R60
Sampling Frequency: 16000Hz
Recording Software: Adobe Audition 1.5

At present in the blind source speech separation algorithm, it was assumed that the number of the mixture signals was equal to or more than the source signals. As we only needed to separate the driver's speech, all the sounds except the driver's were noise. And two microphones could meet the demand of using ICA algorithm. For the speech interface was mainly used to receive the driver's speech, the noise inside vehicle related to the position of the driver. All the microphones used in the noise measurement experiment must install in a certain form. Under the microphone, a thin sponge pad (about 8mm) must be laid to make sure that the vehicle vibration would not influence the microphone. The installation of the microphones should be so tight that they could not make a relative motion for the vehicle. A relative motion meant having amplitude about 20mm.

The vertical distance from microphone to its fixed position on vehicle must be longer than $0.15m$. The two microphones must point to the driver's mouth in their most sensitive direction [17]. One place of the microphones was the dashboard in front of the driver, and the other was sun visor. The vertical coordinate of the microphone was $0.70 \pm 0.05m$ above the intersection of the seat surface without driver and the backrest surface. The horizontal coordinate of the microphone was the center plane of the seat. In the driver's seat, the left distance between the horizontal coordinate and the seat center plane was $0.20 \pm 0.02m$. The position of the microphones related to the seat and the practical position of the microphones show in figure2 and figure3.

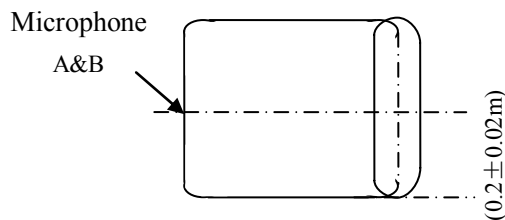
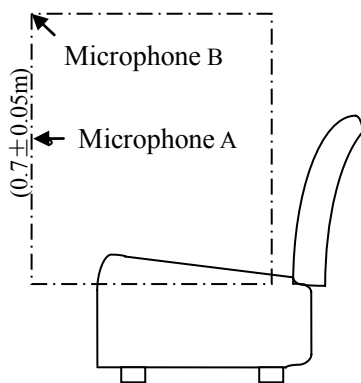


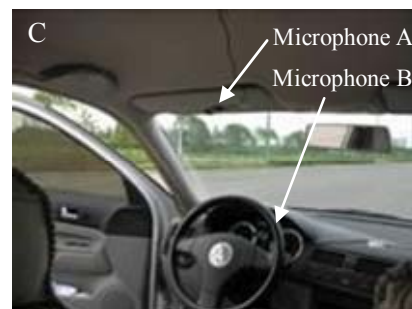
Figure 2. The location of the microphone relative seat



(A) Actual location of microphone A



(B) Actual location of microphone B



(C) Comparison chart of actual location of microphone A and B in the vehicle

Figure 3. Actual location of microphones in the vehicle

In the experiment, according to the specific speech command, a section of words was designed which completely composed of speech commands. Then microphone A recorded a driver's speech commands. Then a selected pure music played in the CD player. Microphone A recorded the CD music. Then microphone A and B recorded the mixture sound at the uniform speed $40km/h$ while the CD played and the driver spoke. The mixture sound was composed of driver's speech, CD music and the noise inside vehicle. The content of the driver's speech was door, window, wiper, CD, air conditioning, direction indicator lamp, and fog light. The music in the CD was the High Mountain and Running River. The sound pressure level in the vehicle was $62.5dB$ with background music, and $69.6dB$ with driver's command, and $67.6dB$ with both. The collected sound stored in the format WAV. And the file length was $4.0s$. The figure 4 and figure 5 were two mixture sounds of driver's speech and CD music. The figure 6 and figure 7 were the separated driver's speech and CD music by FastICA.

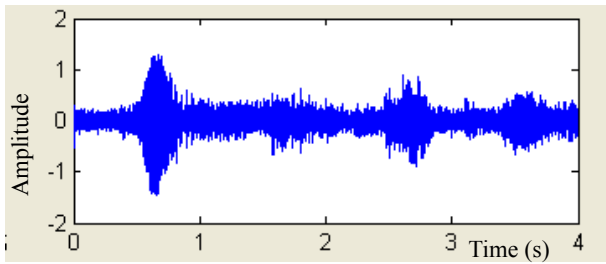


Figure 4. Mixed signals 1

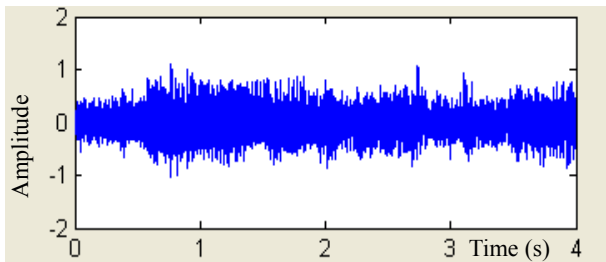


Figure 5. Mixed signals 2

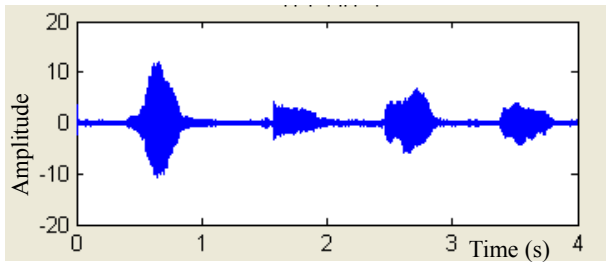


Figure 6. Isolated driver voice

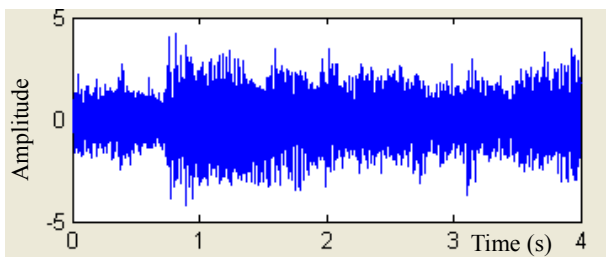


Figure 7. Isolated music of CD

In the four figures above, they showed that the time domain of the signals did not change before and after separation. But the amplitude of the signals increased after separation by the FastICA algorithm. Through listening, it was found that the volume of the speech significantly increases. The spectrograms before and after separation were showed in figure 8, 9, 10 and 11.

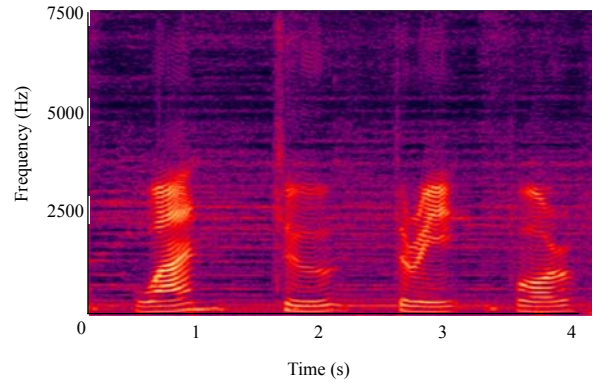


Figure 8. Spectrum language of driver voice

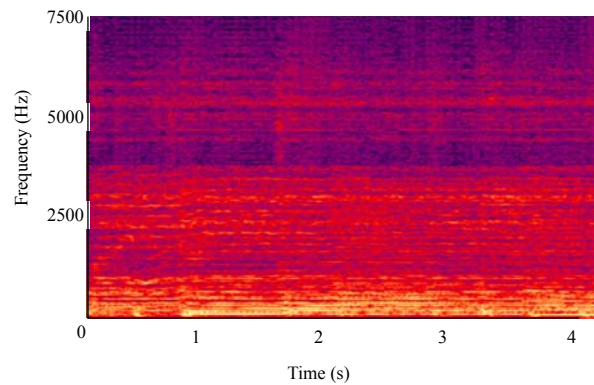


Figure 9. Spectrum language of original CD music

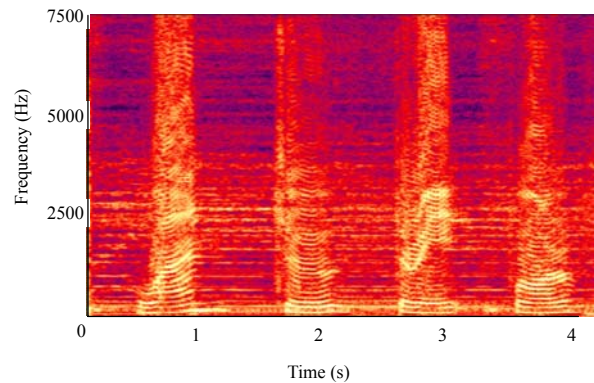


Figure 10. Spectrum language of isolated driver voice

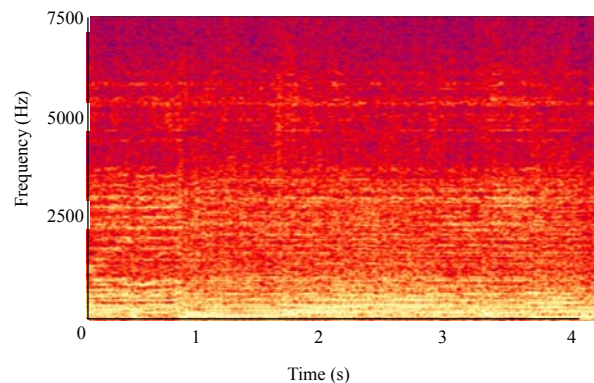


Figure 11. Spectrum language of isolated music of CD

In the four figures, they clearly showed that the FastICA algorithm not only reserved most speech signal energy, but also eliminated the background noise effectively. In figure 10, the reserved sound was driver's speech and the eliminated sounds were noise inside vehicle and CD music. In figure11, the reserved sound was CD music and the eliminated sounds were noise inside the vehicle and the driver's speech. The spectrogram after separation was more similar to the pure original one.

Subject test method was used to test the separated driver's speech and CD music in the experiment. And the evaluation standard was mean opinion score which was widely accepted to test the subject acceptance level [18]. As the driver's speech and the CD music recorded while the vehicle was running, the MOS result of the original driver's speech was 4.39 and the original CD music was 4.51. So the FastICA algorithm was better. The concrete items were in table II.

TABLE II.
THE TEST RESULT OF MOS

Evaluation Group	The original driver's speech	The separated driver's speech by FastICA	The original CD music	The separated CD music by FastICA
1	4.50	3.50	4.60	3.90
2	4.30	3.60	4.60	3.80
3	4.30	3.50	4.50	3.80
4	4.20	3.30	4.30	3.70
5	4.30	3.20	4.50	3.80
6	4.30	3.20	4.50	3.60
7	4.50	3.50	4.50	3.50
8	4.70	3.50	4.50	3.50
9	4.20	3.50	4.50	3.50
10	4.60	3.50	4.60	3.50
Mean	4.39	3.43	4.51	3.58

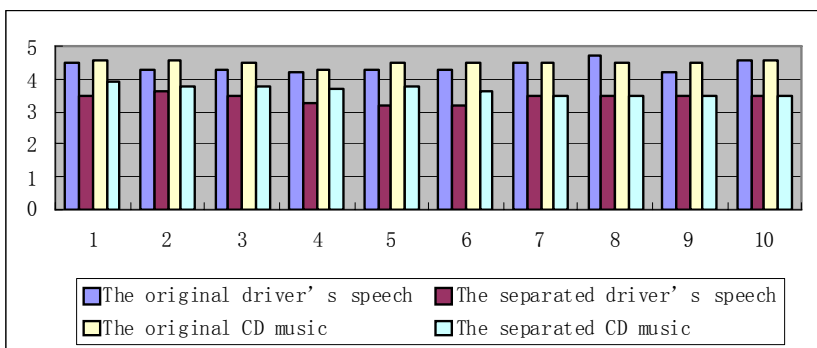


Figure 12. The numerical analysis diagram of the MOS test result

In the MOS test of the experiment, 10 people who never learned BSS or studied the speech signal process were selected as listeners. Then they listened to the

original driver's speech, the original CD music, the separated driver's speech, and the separated CD music. After listening, they scored the test sound by their satisfaction. And the satisfaction level was divided into 5. At last, mean opinion score of every test sound was calculated. Table II was the specific data of the MOS test result. And figure12 was the numerical analysis diagram of the MOS test result.

In the experiment, it showed that the effect of the separated driver's speech and CD music by the FastICA was only understood and barely accepted. It was satisfied in mood and rhythm but not the degree. The score of the original driver's speech and CD music were 4.39 and 4.51, not 5 mainly because the sound were collected when the vehicle was running, which means there were bigger noise interference with the recorded sound.

IV. CONCLUSION

In the paper, the ICA method in the statistics and signal process were applied to the vehicle practical driving environment innovatively. And the method separated the driver's speech successfully and it was the front end process for further recognition of the driver's speech effectively. The reasonable real BORA vehicle experiment validated the performance of the method.

Compared the spectrograms and the Spectrum language before and after separation, and tested by MOS, the FastICA could obtain a better effect in practical vehicle speech separation algorithms. And the method could be the front end process for further recognition.

The next work was to construct a whole high recognition accuracy vehicle speech interface system based on the method in the paper, and the system could meet the demand of the practical application.

ACKNOWLEDGMENT

The National Development & Reforming Committee Foundation of China (No.CNGI-09-01-11), Jilin University basic R&D operating expenses advanced cross-subject innovation projects (No. 200903191) and basic scientific research and innovation projects of young teachers.

REFERENCES

- [1] C. Charlie, Managing the Merger of Computer and Truck Electronics [C]. SAE technical paper, 1999-01-3753.
- [2] V. Cureton. Transforming your operation into a world-class production facility [J], Robotics World, 1999, 17(2), 36-40.
- [3] H. L. Gelgele, K. S. Wang. An expert system for engine fault diagnosis: Development and application [J]. Journal of Intelligent Manufacturing, 1998, 9(6)539-545.
- [4] T. A. Keim. Systems for 42V mass-market automobiles [J]. Journal of Power Sources. 2004, 127(10)16-26.
- [5] Z. Z. Yang, J. Liu, C. Eric, et al. An embedded system for speech recognition and compression [C], 2005 International Symposium on Communications and Information Technologies Proceedings, 2005 II, 287-290.

- [6] Y. Xue, Y. W. A. J. Yang. Milind. Microsoft windows highly intelligent speech recognizer [C]. IEEE International Conference on Acoustics, Speech and Signal Processing, 1995(1) 93-96.
- [7] Z. Alfred, Z. Peter, Psychoacoustic Modelling of Sound Attributes SOUND CHARACTER AND DRIVING NOISE[C], SAE, 2006-01-0098.
- [8] F. Almonte, V. K. Jirsa, E. W. Large, et al. Integration and segregation in auditory streaming, *Physica D: Nonlinear Phenomena*, 2005, 212(1-2), 137-159.
- [9] S. Chiu, C. Lub, D. W. A. C. Wenc. A histogram based data-reducing algorithm for the fixed-point independent component analysis [J]. *Pattern Recognition Letters*. 2007, 29(3): 5.
- [10] I. Kalyakin, N. G. X. Lezb, T. K. X. R. Lyytinenc. Independent component analysis on the mismatch negativity in an uninterrupted sound paradigm [J]. *Journal of Neuroscience Methods*. 2008, 174(2).
- [11] K. Kokkinakis, A. A. K. Nandi. Generalized gamma density-based score functions for fast and flexible ICA [J]. *Signal Processing*. 2006, 87(5).
- [12] J. A. B. Lee, F. X. Vrinsc, A. M. Verleysenc. Blind source separation based on endpoint estimation with application to the MLSP 2006 data competition [J]. *Neurocomputing*. 2007, 72(10).
- [13] ICA Fluor wins Pemex crude oil dehydration contract [J]. *Pump Industry Analyst*. 2008, 2008(9).
- [14] D. Tsai, A. S. Lai. Defect detection in periodically patterned surfaces using independent component analysis [J]. *Pattern Recognition*. 2008, 41(9).
- [15] R. C. Wolf, F. D. Sambatarob, N. Vasic, et al. Aberrant connectivity of lateral prefrontal networks in presymptomatic Huntington's disease [J]. *Experimental Neurology*. 2008, 213(1).
- [16] O. Helcio, L. B. Edgar, M. H. Marcelo. Using SIL/PSIL to Estimate Speech Intelligibility in Vehicles [C]. 2005-01-3973.
- [17] V. Michel. Implementation of a New Metric for Assessing and Optimizing the Speech Intelligibility Inside-Cars [C]. 2005-01-2478.
- [18] I. Kazuya, M. Yonosuke, Y. Noritoshi, et al. Evaluation of a Voice- Activated System Using a Driving Simulator [C]. 2004-01-0232.

Jindong Zhang his research interests mainly cover intelligent control and embedded systems, pattern recognition. He received the ME (2006) and PhD (2009) in computer science and technology of Jilin University. He has more than 20 peer reviewed scientific articles. He is on faculty of Jilin University as a teacher of Computer Science in Jilin University.

Guihe Qin was born in 1962, Prof, PhD Supervisor. He received the BS (1985) in computer engineering from the Dalian University of Technology, ME (1988) in computer engineering and PhD (1997) in communication and electronic systems from the Jilin University of Technology. From 1988 to 2000, he was on the faculty of Jilin University of Technology. He worked in the University of Arizona, US, and KAIST, Korea, as visiting scholar in 1999 and 2001 respectively. He works in the area of real-time and embedded systems. With others or by himself, he has finished more than 30 research projects, and has more than 60 peer reviewed scientific articles and 2 books published. Currently, he is on faculty of Jilin University as a professor of Computer Science and Technology and State key laboratory of automobile dynamic simulation, Jilin University.

Ye Liu was born in 1986. MS.candidate. She received the BS (2009) in College of Computer Science and Technology from Jilin University. Her main research interest is real-time and embedded systems.

Semantic Analysis of Traffic Video Using Image Understanding

Jian Wu^{1,2}

1.The Institute of Intelligent Information Processing and Application, Soochow University, Suzhou 215006, China;

2.Jiangsu Yihe Traffic Engineering Co., Ltd., Suzhou 215002, China

Email: szjianwu@163.com

Zhi-ming Cui^{1,2}, Heng-jun Yue^{1,2}, Guang-ming Zhang^{1,2}

1.The Institute of Intelligent Information Processing and Application, Soochow University, Suzhou 215006, China;

2.Jiangsu Yihe Traffic Engineering Co., Ltd., Suzhou 215002, China

Email: szzmcai@suda.edu.cn

Abstract—Video retrieval technology has always been the important application domain of image engineering. Aiming at the problem that low-level features cannot describe high-level semantic completely and accurately in the content-based video retrieval technology, this paper introduces the methods of image understanding and proposes a video semantic analysis framework using scene analysis for recognizing some semantic events in the traffic video. The experimental results show that the feasibility of the framework is good and the event recognition method based on scene analysis has good recognition effect for vehicle status in traffic video.

Index Terms—video retrieval; semantic analysis; image understanding; scene analysis

I. INTRODUCTION

With the development of society economy and the increasement of living level, vehicle turns into the necessary means of transportation travel. With the increasing number of vehicles, traffic problems are also becoming more serious, traditional television monitoring technology can not meet the requirements of the current traffic management. In recent years, intelligent transportation system (ITS) becomes a new generation of traffic management systems which is based on many new technologies, such as computer network, video transmission, image processing, video processing and computer vision. ITS is applied in many scenarios, including airports, stations, passenger flow, highway intelligent scheduling, operations management, vehicle scheduling. To a certain extent, it improves transportation efficiency, relieves traffic congestion and improves road capabilities. Effective control of the vehicle can reduce traffic accidents, so as to save the social cost. How to extract sensitive data from mass data timely and efficiently has become a research focus.

Most of the existing image retrieval systems describe the image content directly using the traditional low-level image features, such as color, texture, shape. Because of considerable difference between these characteristics and people's understanding, image retrieval based on low-

level features is always unsatisfactory [1~4]. In academic work and application systems, it is not described as an image understanding process. To improve the efficiency as much as possible, most retrieval system process retrieval based on simple image features rather than scene analysis. While browsing the video, we just want to know how much content the video has and watch clips of interest. How to integrate semantic features of images to search retrieval is the key to improve system performance. With the plenty rich content from video analysis and different search requirements, the problem of scene analysis is unable to avoid, and the core problem can be solved only by the image understanding method.

Video image understanding focused on the interpretation of video sequences, both related to the spatial characteristics of images and the time characteristics of the video sequences. This paper focuses on the application of image understanding methods with hierarchical structure in the traffic video semantic analysis. As an important component of traffic incident detection part, the results of vehicle status analysis directly impact on the application level of traffic video analysis technology. According to the real-time and accuracy of video analysis, this paper proposed a new video semantic analysis framework using image understanding. Firstly, obtain the moving vehicles by background subtraction method, then track the detected vehicles with fast normalized cross-correlation method based on prediction to get accurate vehicle trajectory, and extract the interested video information, such as vehicle I/O time, vehicle type, vehicle color, etc. Lastly, based on the above structured data, the useful information for decision can be gotten by video querying and statistical analysis.

II. FRAMEWORK FOR TRAFFIC VIDEO SEMANTIC ANALYSIS

A. Video Image Understanding

Based on the image object, image understanding is a science which takes knowledge for core, researches what the image has, the relationship between them, what scene the image is and how to use the scenarios. Image

understanding, computer vision and artificial intelligence are closely linked. Visual information is the low-level data of image understanding, theoretical starting point is the machine vision; knowledge information is the high-level object of image understanding which based on artificial intelligence [5]. By analyzing and processing of video sequences, the image understanding obtained the understanding scene or object behavior from video. With the progress of technology, more and more applications are partial for the understanding of image sequence and have a wide applications range, including intelligent video surveillance, human-computer interaction, video compression, motion extraction and analysis, robot vision, etc.

Actually, video image understanding is to simulate people to understand visual scene, "visual" means that the human eyes obtain information from the objective world, "sense" means that processing visual information input to the brain. Similar to the human visual process, the purpose of the video image understanding is through the computer to complete the automatic scene understanding [6]. As shown in figure 1, the understanding course of video image understanding has hierarchy.

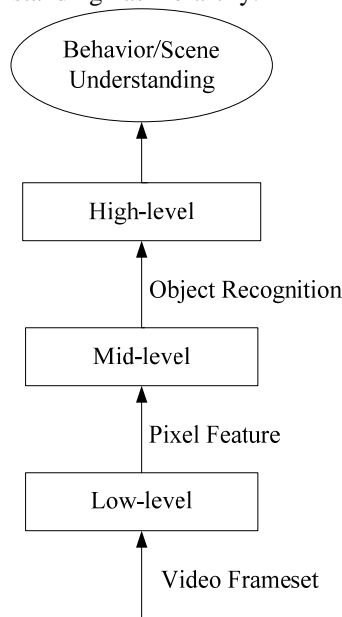


Figure 1. Video image understanding mechanism

The video image segmentation in the lowest level is responsible for separating foreground from background of video sequence, the foreground is concerned content of upper understanding; object recognition and classification in the interlayer is responsible for identifying object in the foreground areas and object classification; the goal of behavior/scene analysis in uppermost layer is to understand the behavior of object and explain the scene state. The specific content at all levels includes:

(1) Video image segmentation.

The processing of input original video frame belongs to pixel-level operations, in fact it uses variety of image

processing technology to transform the image to the image. Although the input image sequence data is great, actually the concern data of the upper understand is very small, always only the moving foreground constantly changing in the sequence, so the video image segmentation must be able to segment the moving target from foreground.

(2) Object recognition.

Object recognition means identification the separated foreground target, understanding their respective categories, such as judge the foreground region is human or pieces of image sequences belonged to the same goal. Similar to the video image segmentation, object recognition and image classification have differences in time-space operations and spatial operations.

(3) Behavior/scene understanding.

Behavior and scene understanding in uppermost layer is responsible for understanding the meaning of action and explaining the state of the scene. Understanding of individual behavior comes from the explanation of motion state and action; in the process of understanding, some prior knowledge is needed inevitably, which can be obtained through prior training. A large number of individual behavior constitute the state of scene, the understanding of scene means the clustering of individual behavior.

B. Application Framework

In the 21st century, intelligent traffic system will be the main trend and it's irreversible. In strengthening the road infrastructure, cities must consider how to improve transport efficiency and safety, reduce resources waste, ensure science and technology for the sustainable development of entire city. With the extensive use of traffic surveillance systems, the amount of traffic data is also growing at an alarming rate. It is unrealistic to manage the video artificially. Although the machine control is superior than people, the primary level of video information intelligent analysis directly leads low utilization efficiency of video information.

Intelligent Transportation System is an integrated transport and management systems, which uses advanced information technology, communication technology, sensor technology, control technology and computer technology, etc. It works in real-time, accurately and efficiently. By close cooperation of people, vehicles and road, to improve transport efficiency, reduce traffic congestion and accidents, alleviate energy consumption and environmental pollution. Some well-known domestic and foreign research institutions and companies are focusing on intelligent transportation technology research and development, and the specific directions are Internet technology, intelligent video analysis, intelligent transportation platforms, etc.

This paper researches traffic video semantic analysis, and its intelligent transportation system framework is shown in figure 2.

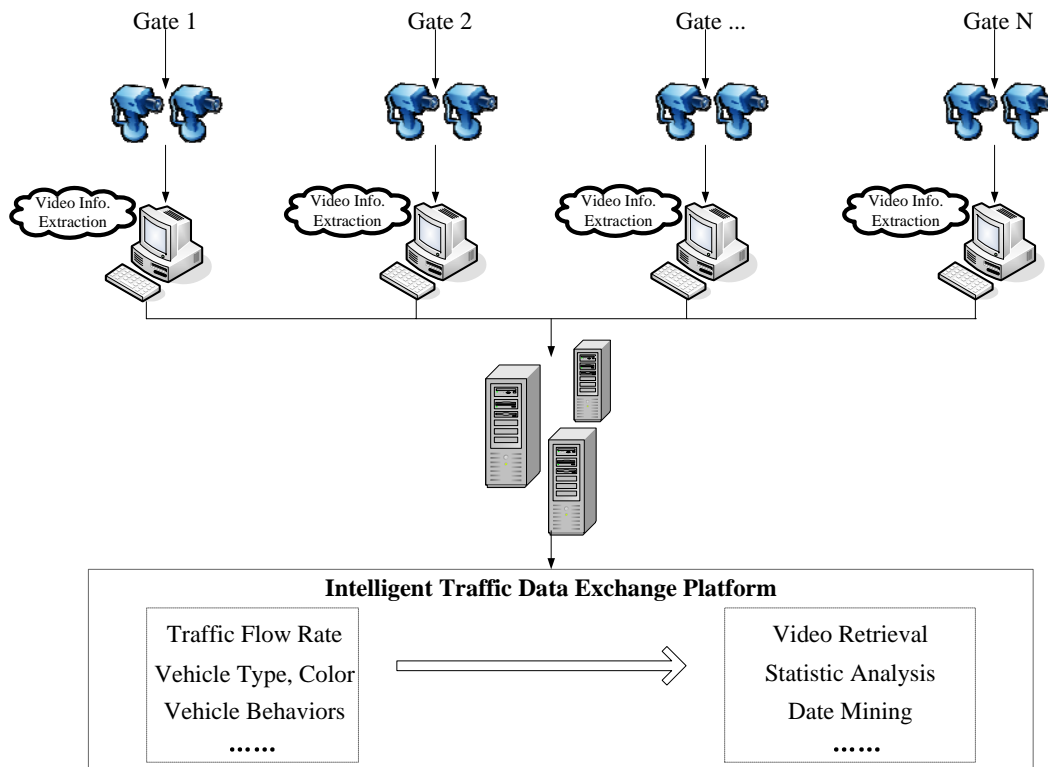


Figure 2. Framework of video analysis technology

Intelligent video analysis technology can be deployed on gate of road security system. After collecting traffic video data in real-time, extract sensitive data traffic from video by the video semantic analysis technology, and send to the server in symbolic form. Then, based on the extracted semantic information, the variety of application systems, using artificial intelligence theory, tools and techniques for statistical analysis and data mining, can serve the urban transportation and management.

III. VEHICLE TRACKING

A. Fast Normalized Cross-correlation

Cross-correlation has some advantages, simple, strong ability of anti-noise, etc. It is a statistical approximation method, commonly used in template matching and pattern

$$\delta(x, y) = \frac{\sum_{i=-m/2}^{m/2} \sum_{j=-n/2}^{n/2} f(x+i, y+j) \cdot t(i, j) - m \cdot n \cdot \mu_f \cdot \mu_t}{\left\{ \left(\sum_{i=-m/2}^{m/2} \sum_{j=-n/2}^{n/2} f^2(x+i, y+j) - m \cdot n \cdot \mu_f^2 \right) \cdot \left(\sum_{i=-m/2}^{m/2} \sum_{j=-n/2}^{n/2} t^2(i, j) - m \cdot n \cdot \mu_t^2 \right) \right\}^{1/2}} \quad (2)$$

In the formula (2), $\delta(x, y)$ is the NCC coefficient of the image f and template t , while m and n is the size of the template t , while μ_f and μ_t are respectively calculate by the formula (3) and formula (4).

$$\mu_f(x, y) = \frac{1}{m \cdot n} \sum_{i=-m/2}^{m/2} \sum_{j=-n/2}^{n/2} f(x+i, y+j) \quad (3)$$

recognition. Get the image f and the template t , to measure the similarity c :

$$c(u, v) = \sum_{x, y} f(x, y) t(x - u, y - v) \quad (1)$$

But there is some shortages in the process of matching using the formula (1). For example, when the grayscale of the image varies with the position, the correlation of the matching area of the template and image may smaller than the correlation of some bright spot in the template and image, resulting in failure of image matching. Besides, the value of $c(u, v)$ is dependent on the size of the template. To solve this problem, it is necessary to normalize the cross-correlation. The coefficient of the so-called normalized cross correlation (NCC) is shown as formula (2).

$$\mu_t(x, y) = \frac{1}{m \cdot n} \sum_{i=-m/2}^{m/2} \sum_{j=-n/2}^{n/2} t(i, j) \quad (4)$$

It can be seen from the formula of NCC, the more resemble the template and the area to be matched are, the more approach to the value of $\delta(x, y)$ is, and it can effectively solve the existing problem in matching of cross-correlation. But it can be seen in the same time that the process of cross-correlation normalizing involve many operation of square root and involution, resulting in a lot

of calculation, and the complexity of the algorithm depends on the size of template t . So in this article we use fast normalized cross-correlation algorithm (FNCC) [7][8], reducing the complexity of the traditional normalized cross-correlation algorithm.

FNCC make use of the concept of sum-table proposed by J.P.Lewis, improving the formula of NCC using sum-table [9]. When $x=0,1,2,\dots,M-1$, and $y=0,1,2,\dots,N-1$, the sum-table of a two-dimensional discrete variable $f(x, y)$ is expressed as:

$$S_{\mu}(x, y) = f(x, y) + S_{\mu}(x-1, y) + S_{\mu}(x, y-1) - S_{\mu}(x-1, y-1) \quad (5)$$

$$\sum_{i=-m/2}^{m/2} \sum_{j=-n/2}^{n/2} f(x+i, y+j) = S_{\mu}(x+m/2, y+n/2) - S_{\mu}(x-m/2-1, y+n/2) -$$

$$S_{\mu}(x+m/2, y-n/2-1) + S_{\mu}(x-m/2-1, y-n/2-1) \quad (8)$$

$$\sum_{i=-m/2}^{m/2} \sum_{j=-n/2}^{n/2} f^2(x+i, y+j) = S_{\sigma}(x+m/2, y+n/2) - S_{\sigma}(x-m/2-1, y+n/2) -$$

$$S_{\sigma}(x+m/2, y-n/2-1) + S_{\sigma}(x-m/2-1, y-n/2-1) \quad (9)$$

$$\sum_{i=-m/2}^{m/2} \sum_{j=-n/2}^{n/2} f(x+i, y+j) \cdot t(i, j) = S_c(x+m/2, y+n/2) - S_c(x-m/2-1, y+n/2) -$$

$$S_c(x+m/2, y-n/2-1) + S_c(x-m/2-1, y-n/2-1) \quad (10)$$

We substitute the above three formula in the process of cross-correlation normalizing, namely FNCC. Then we can introduce the FNCC method to the object tracking, and update the moving object basing on forecasting.

B. Vehicle Tracking Based on Prediction

In this paper, we update the tracking template in real time [10][11], making use of the moving vehicle tracked in the current frame to update the tracking template. There is a feature that the moving vehicle is stiff object, so the vehicles in two adjacent frames changes little. We can track the vehicles make use of it. To some extent, it meets the requirement of high matching probability of normalized cross-correlation. So it improves the efficiency of tracking as well as the accuracy of tracking.

We can get the moving trajectory of tracked object in the tracking process, and predict the position of the moving object in the next frame [12]. Accordingly, we should normalize the moving templates and the predicted regions fast, and it can reduce the time consumed in the fast normalized cross-correlation. In the tracking process, at first we should track the moving object and then get the trajectory of it. Suppose the position of the moving vehicle in the current frame f^n is P_n while the position in the former frame f^{n-1} is P_{n-1} . Then the size of the area to be matched refers to the area position of the moving object in the current frame. We can calculate it in accordance with the formula (11) and formula (12) as follows:

$$SR.x = t.x + 2 \times (P_n.x - P_{n-1}.x) \quad (11)$$

$$SR.y = t.y + 2 \times (P_n.y - P_{n-1}.y) \quad (12)$$

$$S_c(x, y) = f^2(x, y) + S_c(x-1, y) + S_c(x, y-1) - S_c(x-1, y-1) \quad (6)$$

$$S(x, y) = f(x, y) \cdot t(x, y) + S(x-1, y) + S(x, y-1) - S(x-1, y-1) \quad (7)$$

In the formula (5), (6) and (7), if $x < 0$ or $y < 0$, $S_{\mu}(x, y)$, $S_{\sigma}(x, y)$ and $S_c(x, y)$ are all equal to 0. We can reform the NCC formula by the three sum-table. In

the formula (2), $\sum_{i=-m/2}^{m/2} \sum_{j=-n/2}^{n/2} f(x+i, y+j)$, $\sum_{i=-m/2}^{m/2} \sum_{j=-n/2}^{n/2} f^2(x+i, y+j)$ and $\sum_{i=-m/2}^{m/2} \sum_{j=-n/2}^{n/2} f(x+i, y+j) \cdot t(x+i, y+j)$ are shown as formula (8), formula (9) and formula (10):

In the formula (11) and formula (12), $SR.x$ and $SR.y$ is respectively the length and width of the predicted area, and $t.x$ and $t.y$ is respectively the length and width of the template, while $P_n.x$ and $P_n.y$ are the adjacent points in the trajectory, as well as $P_{n-1}.x$ and $P_{n-1}.y$.

IV. VIDEO CONTENT EXTRACTION

A. Vehicle Type Recognition

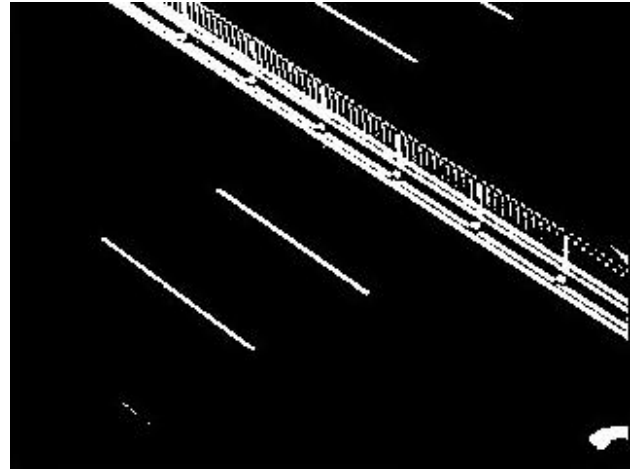
Vehicle type identification is an important component of the Intelligent Transportation system. It is now the hot spot domestic and foreign. This paper presents a method of vehicle type identification based on drive line detection and divided the vehicles into four types, they are car, truck, van and bus.

Basic steps of vehicle type identification are as follows: (1) First get the background image of the current video frame.

Background subtraction method is faster than the inter-frame subtraction method and optical flow method, and can detect the moving object accurately. So we adopt the method to get the moving regions. Before the operation of subtracting background, we should first build the background model. It is reasonable to assume that for some time there is only small gray scale change in the background while the gray scale of the prospect varies a lot to different vehicles, even the different parts of the same car is not the same. The schematic diagram of background extraction course is shown in figure 3.



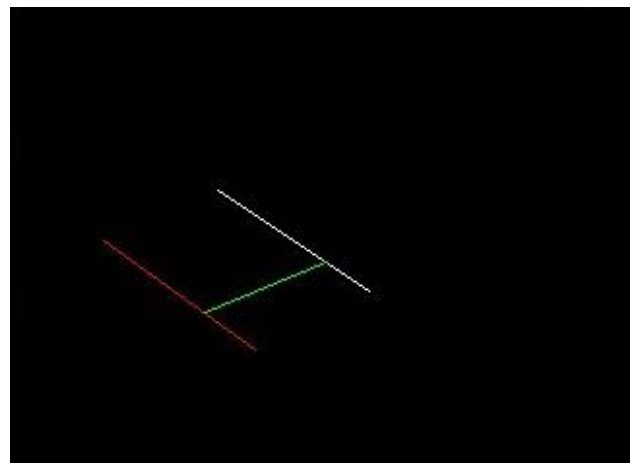
(a) Video Frame



(a) Binarization of background frame



(b) Extracted background frame



(b) Calibration of drive line

Figure 3. Background extraction

(2) Identify two drive lines using Hough Transform.

It is often necessary in computer vision to identify or locate some geometry, such as lines, circles, ellipses, and other graphics. Hough transform provides a method to detect the lines and is simple. Which is developed to detect circles, ellipses, and other general graphics, Hough transform's theory is to map a set of points of some graphic to one point. The point record the number of the points in the set and the number recorded assist the program in finding the point. The point is the parameter of the graphic and the range of the parameter is called parameter space. Hough transform using straight lines with polar coordinates to represent. There is two parameters and they are associated by polar-style, so the voting process only need to traverse one of them and search the peak in the two-dimensional parameter space. The schematic diagram of drive line width determination is shown in figure 4.

Figure 4. Determination of drive line width

(3) Using the ratio of measured drive line width and actual width to calculate actual area of vehicle S_{Real} .

$$S_{Real} = \left(\frac{d}{l}\right)^2 S_{Oval} \tag{13}$$

$$S_{Oval} = \pi ab \tag{14}$$

Where d is the actual lane width, l the measured lane width, a and b are the lengths of the major semi-axis and the minor semi-axis of the external ellipse.

(4) Classify the vehicles with the actual area value of the vehicle S_{Real} .

B. Color Recognition

First we should extract the color information and classify it to the specified color category. There are several key questions. First, we should select the appropriate color space. Second is to define the color characteristics and quantified. At last, we must measure the similarity and match, in other words, how to define the similarity of the color features and match fast. As the RGB color space is used universally in the images and it is easy to access the RGB color of the vehicles. But there is a problem that we can't commonly use Euclidean

distance to properly describe the distance of two different colors.

In addition, when the color changes the three value of R, G and B will have non-linear change, it is not conducive to the color quantization and classification. Therefore, we choose HSL space in this paper and we will first transform the RGB color space into the HSL color space.

Make \max the maximal value of R, G and B, while \min the minimal. Then the formula used to transform the RGB color space into the HSL color space is as follow :

$$H = \begin{cases} 0^\circ, & \text{if } \max = \min \\ 60^\circ \times \frac{G - B}{\max - \min} + 0^\circ, & \text{if } \max = R \text{ and } G \geq B \\ 60^\circ \times \frac{G - B}{\max - \min} + 360^\circ, & \text{if } \max = R \text{ and } G < B \\ 60^\circ \times \frac{B - R}{\max - \min} + 120^\circ, & \text{if } \max = G \\ 60^\circ \times \frac{R - G}{\max - \min} + 240^\circ, & \text{if } \max = B \end{cases} \quad (15)$$

$$L = \frac{1}{2}(\max + \min) \quad (16)$$

$$S = \begin{cases} 0 & \text{if } L = 0 \text{ or } \max = \min \\ \frac{\max - \min}{\max + \min} = \frac{\max - \min}{2L}, & \text{if } 0 < L \leq \frac{1}{2} \\ \frac{\max - \min}{2 - (\max + \min)} = \frac{\max - \min}{2 - 2L}, & \text{if } L > \frac{1}{2} \end{cases} \quad (17)$$

The analysis found that L component to a large extent dictate the colors of black, white and gray, in other words, when the L component is in a fixed range the color will be one of the three colors no matter what the value of H component and S component is. So the method to quantify and classify the color is to distinguish the colors of black, white and gray according to the value of L component and then determine the other colors according to the value of H component. With this method we can divide the HSL color space into seven colors of black, white, gray, red, yellow, green and blue, which comprise almost all the vehicles of pure color in the video.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. The Tracking Effect

The screenshots of this article is from the video which is shot in the tunnel of DuShuHu in Suzhou and processed with the tracking algorithm of normalized cross-correlation. The tracking result is shown in figure 5 and figure 6. The fifth vehicle drive in the monitoring area at 624th frame and out at the 657th frame. The fifth figure intercepts the 630th frame and 650th frame, from which we can see the currently tracking although the closer the vehicle away from, the bigger it is. Similarly, the two frames intercepted show that the van drive in at 931st frame and out at 960th frame. Although the size

and shape of the van changes, we can calibrate the van correctly and track it in real-time.



(a) One moment of vehicle 5# tracking

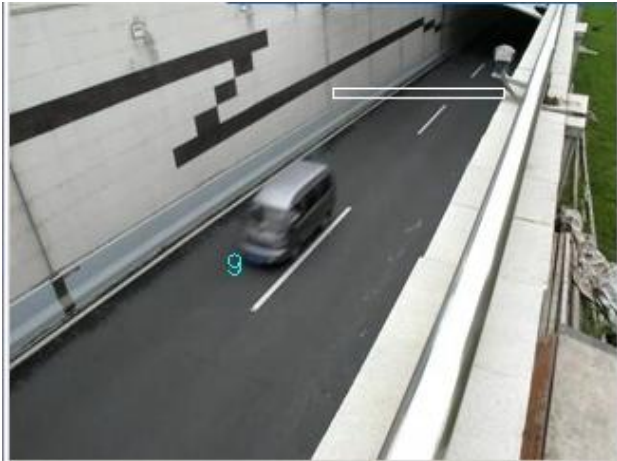


(b) Another moment of vehicle 5# tracking

Figure 5. The tracking course of vehicle 5#



(a) One moment of vehicle 9# tracking



(b) Another moment of vehicle 9# tracking

Figure 6. The tracking course of vehicle 9#

B. Video Content Extraction Result

Based on accurately tracking, we can extract the video information we are interested in with the method of vehicle style identification and color classification designed above. In this paper, we primarily judge and find the numbers of the vehicles, the frame number the vehicles drive in and out, the vehicle category and the vehicle color, and eventually write in the database for the further experiments. It is shown in Figure 7 that the algorithm involved in this article can identify the data of the vehicle mentioned above and finally write them to the database.

cameraID	videoID	vehicleID	frameIn	frameOut	vehicleType	vehicleColor	memo1
0101	2009072806	1	167	184	Van	White	
0101	2009072806	2	661	703	Car	Black	
0101	2009072806	3	719	726	Car	White	
0101	2009072806	4	924	957	Car	Black	
0101	2009072806	5	1042	1042	Car	Black	
0101	2009072806	6	1066	1066	Car	Silver	
0101	2009072806	7	1231	1260	Van	Blue	
0101	2009072806	8	1745	1772	Van	Silver	
0101	2009072806	9	2094	2103	Car	Blue	
0101	2009072806	10	2339	2377	Car	Blue	
0101	2009072806	11	2419	2448	Car	Black	
0101	2009072806	12	2614	2657	Car	Black	
0101	2009072806	13	2623	2623	Car	White	
0101	2009072806	14	2673	2708	Car	Black	
0101	2009072806	15	2940	2940	Car	White	
0101	2009072806	16	3144	3144	Car	Yellow	
0101	2009072806	17	3172	3342	Van	Black	
0101	2009072806	18	3352	3352	Car	Black	
0101	2009072806	19	3470	3497	Car	Yellow	
0101	2009072806	20	3539	3566	Car	Yellow	
0101	2009072806	21	3741	3748	Van	White	
0101	2009072806	22	3857	3889	Car	Black	
0101	2009072806	23	3892	3892	Car	White	
0101	2009072806	24	4642	4665	Van	White	
0101	2009072806	25	4929	4929	Car	White	
0101	2009072806	26	5009	5037	Car	Black	
0101	2009072806	27	5013	5016	Car	Black	

Figure 7. Extraction result of video content

C. Video Querying

After the above steps, we transform the video data into text and can query them for the specific needs. As is shown in the figure 8, the users set the search criteria,

including the date, the time, vehicle style, vehicle color and so on. According information written in the previous steps, the system will return back the frame set and extract the corresponding video frames in real-time. So the users can quickly check the video clip they want.

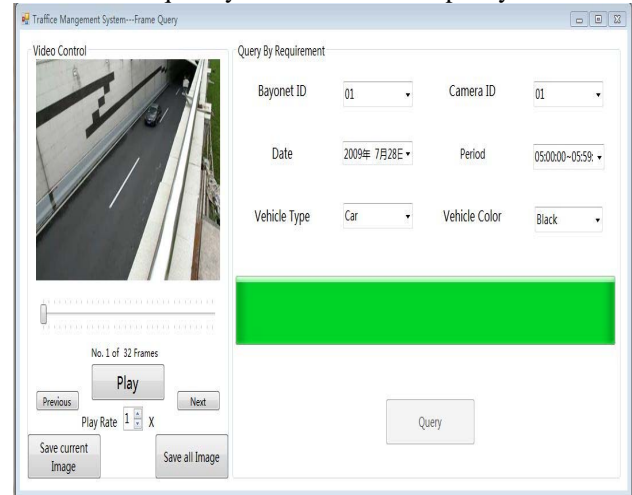
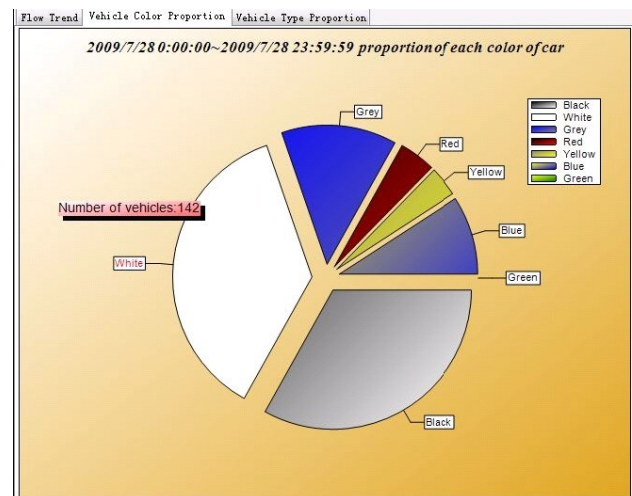


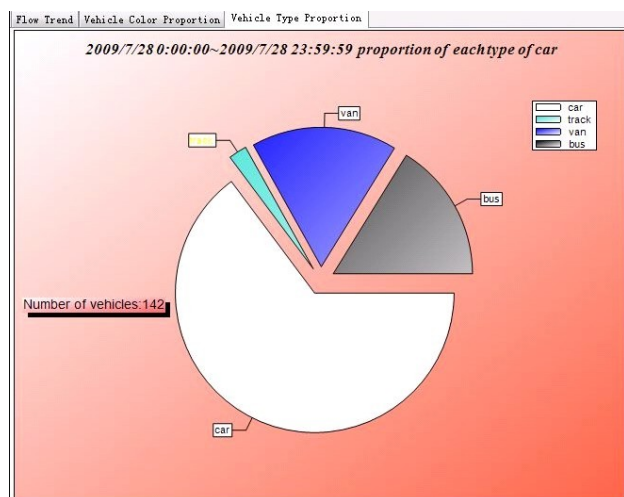
Figure 8. Video querying window

D. Statistical Analysis

With the methods involved in the paper, users are allowed to set a specific date and time ,and statistic the vehicle style and color, finally get the percents of all the categories and colors. The statistic of the vehicle category can show the current road situation, while the statistic of the vehicle color can show the preference vehicle color of the public and assist the manufacture and trade company in making a strategic decision. These statistics can also assist the traffic control department with the corresponding analysis and decision. As is shown, the figure 9(a) is the ratio diagram of the vehicle colors of the video got in one day of July in 2009, while the figure 9(b) is the ratio diagram of the vehicle types.



(a) According to vehicle color



(b) According to vehicle type

Figure 9. Vehicle statistical analysis

VI. CONCLUSION

Regarding to the problem that traditional video retrieval methods can't make full use of semantic information of video flow, this paper introduces the image understanding method to traffic video analysis, and proposes a video semantic analysis framework based on scene analysis. Firstly, obtain the moving vehicles by background subtraction method, then track the detected vehicles with fast normalized cross-correlation method based on prediction to get accurate vehicle trajectory, and extract the interested video information, such as vehicle I/O time, vehicle type, vehicle color, etc. Lastly, based on the above structured data, the useful information for decision can be gotten by video querying and statistical analysis. The experimental results verified the feasibility of the framework.

ACKNOWLEDGEMENT

This research was partially supported by the Natural Science Foundation of China under grant No. 60970015, the 2009 Special Guiding Fund Project of Jiangsu Modern Service Industry (Software Industry) under grant No. [2009]332-64, the 2010 Program for Postgraduates Research Innovation in University of Jiangsu Province under grant No. CX10B_041Z, the Applied Basic Research Project (Industry) of Suzhou City under grant No. SYJG0927 and No. SYG201032, and the Beforehand Research Foundation of Soochow University.

REFERENCES

- [1] Lee D, Barber R, NiBlack W, et al. Indexing for complex queries on a query-by-content image database[C]. In Proceedings of ICPR 1994, pp.142-146.
- [2] Pentland A, Picard R W, Sclaroff S. Photobook: content-based manipulation of image database[J]. International Journal of Computer Vision, 1996, 18(3):233-254.
- [3] Simith J R, Chang S F. Tools and techniques for color image retrieval[C]. In Proceedings of IS&T/SPIE 1996, vol.2670, pp.1-12.
- [4] Wu J K. Content-based indexing of multimedia database[J]. IEEE Transactions on Knowledge and Data Engineering, 1997, 9(6):978-989.
- [5] XIE Zhao. Research for key issues and methods in image understanding[D]. Hefei University of Technoly, 2007.
- [6] LIANG Ying-hong, WANG Zhi-yan, CAO Xiao-ye, XU Xiao-wei. Application of video image understanding to measurement of pedestrain flow[J]. Computer Engineering and Design, 2008, 29(5):1203-1206.
- [7] D. M. Tsai, C. T. Lin. Fast normalized cross-correlation for defect detection[J]. Pattern Recognition Letters, 2003, 24, 2625-2631.
- [8] GUO Wei, ZHAO Yi-Gong, XIE Zhen-Hua. An Improved Normalized Cross-correlation for Template Matching of Infrared Image[J]. Acta Photonica Sinica, 2009, 38(1) : 189-193.
- [9] Lewis J P. Fast normalized cross-correlation [A]. Proceeding of Vision Interface[C]. Canada Quebec, 1995, 120-123.
- [10] HOU Zhi-Qiang, HAN Chong-Zhao. A Survey of Visual Tracking[J]. Acta Automatica Sinica, 2006, 32(4), 603-617.
- [11] XUE Chen, ZHU Ming, LIU Chun-xiang. Review of tracking algorithms under occlusions[J]. Chinese Journal of Optics and Applied Optics, 2009, 2(5), 388-393.
- [12] Jian Wu, Heng-jun Yue, Zhi-ming Cui, Jian-ming Chen. Moving Object Tracking Method Based on Adaptive On-line Clustering and Prediction-based Cross-correlation[C]. In Proceedings of IEEE CIT 2010, Bradford, UK, pp.487-494.



Jian Wu was born in Nantong on the 29th April, 1979, and got master degree in the field of computer application technology from Soochow university, Suzhou city, China in 2004. The main research direction is computer vision, image processing and pattern recognition.

He works as a teacher in the same college after his master graduation. Now he is pursuing the doctoral degree. He was awarded the Third Prize of 2007 Suzhou City Science and Technology Progress and the 2008-2009 Soochow University Graduate Scholarship Model.

Zhi-ming Cui was born in Shanghai on the 4th July, 1961. Professor, PhD Candidate Supervisor. The main research direction is deep web and video mining.

Heng-jun Yue was born in Henan on the 11th July, 1984. Master, his main research direction is video processing.

Guang-ming Zhang was born in Suzhou on the 10th Februry, 1981. PhD Candidate, his main research direction is image processing and video retrieving.

Research on Video Quality Assessment

Chunting Yang

Zhejiang University of Science and Technology, China

Email: hzyangct@gmail.com

Yang Liu and Jing Yu

Zhejiang University of Science and Technology, China

University of Toronto, Canada

Email: hzliuyang@gmail.com, jyu@mie.utoronto

Abstract—Objective quality assessment plays a very important role in the video applications, as they promise the means to evaluate the performance of acquisition, display, coding and communication systems. Objective video quality assessment still has a long way to go before it reaches the level of success. Perception video sequences quality metrics are of great potential benefit to the video industry. Many researchers have focused on developing digital video sequences quality metrics which produce results that accurately emulate subjective responses. However, to be widely applicable a metric must also work over a wide range of quality, and be useful for in-service quality monitoring. In this paper, we propose novel quality metrics for video sequences. The temporal correlations of video frames and the visual interest feature are considered in this method. Meanwhile the metrics are capable of capturing spatial distortions in video sequences. And they can quantify the spatial distortions and differentiate the type of distortion. Furthermore the metrics correlate well with the subjective quality measures because perception distortions of human have been taken into consideration. Experimental results show that our perceptual quality metrics performs better than the existing methods.

Index Terms—video quality assessment; perception distortions; motion intensity

I. INTRODUCTION

Digital video was first introduced commercially in 1986 with the Sony. Consumer digital video first appeared in the form of QuickTime, Apple Computer's architecture for time-based and streaming data formats, which appeared in crude form around 1990. Now a digital video has pervaded the lives of people due to the popularity of applications such as Internet Video, Interactive Video on Demand (VoD), Video Phones, Personal Digital Assistant (PDA) and other Wireless Video devices, Video Surveillance, HDTV, Digital Cinema etc. Visual content is still one of the most challenging content types in content distribution infrastructures. It is not only challenging in terms of bandwidth and timing requirements but also because of the user quality perception. With the rapid development of video technology, how to effectively evaluate the quality of video sequences has become a research hot topic. A great deal of effort has been made to develop novel objective image and video quality assessment methods [16][17][18]. The content distribution has to support a large number of

end users. The most reliable way of assessing the quality of an image or video is subjective evaluation, because human beings are the ultimate receivers in most applications. The mean opinion score (MOS), which is a subjective quality measurement obtained from a number of human observers, has been regarded for many years as the most reliable form of quality measurement. However, the MOS method is very tedious, expensive and impossible to be executed automatically for most applications. Instead, an objective image or video quality metric can provide a quality value for a given image or video automatically in a relatively short time. This is very important for real world applications.

All types of objective video quality assessments are based on measurement of differences between the original and degraded video sequences in some way. Conventional, objective video quality assessment methods are based on statistics features[4,6,8]. The absolute difference between the reference and degraded video sequences was calculated for each pixel. These absolute differences are then transformed into a quality score by using various statistical methods. The most popular video quality metrics based on statistical processing of absolute pixel differences are MSE (Mean Square Error), SNR (Signal-to-Noise Ratio), PSNR (Peak Signal-to-Noise Ratio), CZD (Czekanowski distance), and Minkowski Distance. The main disadvantage of these video quality metrics is that they do not correlate well with subjective quality measures because perception distortions of human have not been taken account of. Recently, several efforts were made to develop metrics that would correlate better with subjective assessment. Metrics like SSIM and VQM are based on the properties of the Human Visual System (HVS) and instead of directly comparing pixel values the metrics subtract structural information from the video frames and compare it. Their predictions are more reliable than those of pixel difference based methods. Over the time several tests[1][5] were performed to find the best metric. The tests showed that most of the pixel-difference methods had statistically equivalent performance and only a few metrics like SSIM and VQM, had a statistically better performance than PSNR.

A great deal of effort has been made to develop objective video quality assessment methods, which incorporate perceptual quality measures by considering HVS characteristics. Some of the developed models are commercially available. However, image and video quality assessment is still far from being a mature research

topic. In fact, only limited success has been reported from evaluations of sophisticated HVS-based quality assessment models under strict testing conditions and a broad range of distortion and image types. Studies conducted by the Video Quality Experts Group indicate that the performance of HVS-based VQA algorithms leaves considerable room for improvement [8].

The assessments of video sequences quality usually are based on the single frame quality, and the motion information inter frames is not considered. Therefore the results cannot be a good match with the actual subjective quality. Research revealed that Human visual system is more concerned about the movement of objects in the scene, and may ignore most of the background information. Some perception factors that have significant effect on single-frame image quality, such as blocking, have little effect on the quality of the entire video sequence. Therefore the assessment of video sequences quality must consider the motion information, in addition to using single-frame image quality.

In this paper, research has been focused on developing novel objective evaluation metrics which enable prediction of the perceived quality level. However, designing objective quality metrics is very difficult due to the limited understanding of the HVS. It is believed that effective objective quality metrics are feasible when the temporal correlations of video frames and the visual interest feature are considered. Meanwhile the knowledge about the distortion types is important for spatial information. Therefore the metrics will be capable of assessing the video sequences quality according to motion intensity. And it can quantify the spatial distortion and differentiate the type of distortion. Experimental results show that the video metrics are reliable, and is well agreed with perceived quality.

II. SPATIAL DISTORTION METRIC

It is believed that the perceived quality of the video after be compressed is often a function of the input scene. That means the perceived video quality is sensitive to the spatial information. A scene which contains a large amount of spatial detail will appear quite distorted at the same bit rate. Objective measurements could quantify the perceived spatial distortions in a way that correlates as closely as possible with the response of a HVS. The difficulty in compressing a given video sequence depends upon the perceived spatial information presented in that video sequence. Perceived spatial information is the amount of spatial detail in the video scene that is perceived by the viewer. Thus, it would be very useful to have approximate measures of perceived spatial information.

Most of the state of the art image and video coding standards, such as JPEG, H.26x and MPEG-1/2/4, make use of the block-based discrete cosine transform. One of the most noticeable distortions is the "blocking distortion". These blocking distortion are structural disturbance, and are sometimes "buried" in the massively accumulated across-the-board pixel-wise error.

A. Edge information

HVS is sensitive to structural information, which is mainly presented by edges, in scene. It is important to calculate the edges energy using spatial information.

Therefore the spatial information is obtained when the video frame is filtered. At time t , the gradient vector $g(m, n, t)$ over the pixels in each filtered frame is then computed as.

$$g(m, n, t) = \left\{ [g_h(m, n, t)]^2 + [g_v(m, n, t)]^2 \right\}^{1/2} \quad (1)$$

$$\varphi(m, n, t) = \arctg \left[\frac{g_v(m, n, t)}{g_h(m, n, t)} \right]. \quad (2)$$

where $g_h(m, n, t)$, $g_v(m, n, t)$ and $\varphi(m, n, t)$ are the horizontal vector, the vertical vector and the angle of $g(m, n, t)$ at the (m, n) in the t^{th} frame in the video sequence, respectively.

B. Spatial detail information

The amount of spatial detail can be presented by standard deviation of the gradient image. Thus, the σ_x and σ_y can be used to measure the perceived spatial information approximately as.

$$\mu_x = \frac{1}{N} \sum_{i=1}^N x_i. \quad (3)$$

$$\mu_y = \frac{1}{N} \sum_{i=1}^N y_i. \quad (4)$$

$$\sigma_x = \left[\frac{1}{(N-1)} \sum_{i=1}^N (x_i - \mu_x)^2 \right]^{1/2} \quad (5)$$

$$\sigma_y = \left[\frac{1}{(N-1)} \sum_{i=1}^N (y_i - \mu_y)^2 \right]^{1/2}. \quad (6)$$

where x_i and y_i are the pixel values of the gradient image x and the degraded gradient image y , respectively.

The distortion of the spatial information $c(x, y)$ is defined as

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_1}{\sigma_x^2 + \sigma_y^2 + C_2}. \quad (7)$$

where $C_1 = (k_1L)^2$ and $C_2 = (k_2L)^2$, and $k_1 = k_2 = 0.03$. L is the dynamic range of pixel values. They were only used when both σ_x and σ_y are close to zero.

There is no distortion of spatial information when $c(x, y)$ equals to 1. This operation is repeated for each frame in the video sequence and results in a time series of spatial information values. Thus, the distortion of spatial information for a video could be obtained by calculating the mean of $c(x, y)$ for each frame in a video sequence.

C. Orientation distortion spatial information

There exists orientation selectivity in HVS. Research shows that HVS is more sensitive to horizontal and vertical distortions than others. Thus horizontal and vertical distortions should be separated from diagonal distortions. The distortions correlate well with perception distortions of human.

If μ_{xhv} is the mean of horizontal and vertical gradient values, and μ_{xnhv} is the mean of diagonal gradient values. x and y are respectively the original and the degraded video. Then the orientation distortion can be defined as:

$$L(x, y) = \frac{\mu_{xhv} / \mu_{xnhv}}{\mu_{yhv} / \mu_{ynhv}} \tag{8}$$

The condition that $L(x, y) < 1$ indicates that the horizontal and vertical edges of the degraded frame are enhanced undeservedly. This kind of distortion is generated by blocking distortion. The worse the blocking distortion is, the smaller the value of $L(x, y)$.

Otherwise $L(x, y) > 1$ indicates that the horizontal and vertical blur distortion occurs. Larger $L(x, y)$ value depicts worse horizontal and vertical blur distortion

The sensitivity of the HVS to the errors may be different for different types of errors. $c(x, y)$ and $L(x, y)$ were expected to quantify the strength of the distortion between the original and the degraded video in a perceptually meaningful way.

III. PERCEPTUAL TEMPORAL INFORMATION

Research revealed that Human visual system is more concerned about the movement of objects in the scene, and may ignore most of the background information. Some perception factors that have significant effect on single-frame image quality have little effect on the quality of the entire video sequence. Therefore the assessment of video sequences quality must consider the motion information, in addition to using single-frame quality.

A. Motion estimation

Motion estimation is the process of determining motion vectors that describe the transformation from one 2D image to another, usually from adjacent frames in a video sequence. It is an ill-posed problem as the motion is in three dimensions but the images are a projection of the 3D scene onto a 2D plane. The motion vectors may relate to the whole image or specific parts, such as rectangular blocks, arbitrary shaped patches or even per pixel. The motion vectors may be represented by a translational model or many other models that can approximate the motion of a real video camera, such as rotation and translation in all three dimensions and zoom.

Block Matching (BM) is a way of locating matching blocks in a sequence of video frames for the purposes of motion estimation. The purpose of a block matching algorithm is to find a matching block from a frame i in some other frame j , which may appear before or after i . Typically each macro block (16×16) in the new frame is compared with shifted regions of the same size from the

previous decoded frame, and the shift which results in the minimum error is selected as the best motion vector for that macro block. BM can be very computationally demanding if all shifts of each macro block are analyzed. This is known as exhaustive search BM. Block matching algorithms make use of criteria to determine whether a given block in frame j matches the search block in frame i .

Sum of Absolute Differences (SAD) is a widely used, extremely simple video quality metric used for block-matching in motion estimation. It works by taking the absolute value of the difference between each pixel in the original block and the corresponding pixel in the block being used for comparison. These differences are summed to create a simple metric of block similarity.

$$SAD = \sum_{i=0}^M \sum_{j=0}^N |c_{ij} - p_{ij}| \tag{9}$$

SAD is an extremely fast metric due to its simplicity; it is effectively the simplest possible metric that takes into account every pixel in a block. Therefore it is very effective for a wide motion search of many different blocks. SAD is also easily parallelizable since it analyzes each pixel separately, making it easily implementable with such instructions as MMX and SSE2. Once candidate blocks are found, the final refinement of the motion estimation process is often done with other slower but more accurate metrics, which better take into account human perception. These include the sum of absolute transformed differences (SATD), the sum of squared differences (SSD), and rate-distortion optimization.

Sum of absolute transformed differences (SATD) is a widely used video quality metric used for block-matching in motion estimation. It works by taking a frequency transform, usually a Hadamard transform, of the differences between the pixels in the original block and the corresponding pixels in the block being used for comparison. The transform itself is often of a small block rather than the entire macro block. SATD is much slower than the SAD, both due to its increased complexity and the fact that SAD-specific MMX and SSE2 instructions exist, while there are no such instructions for SATD. However, SATD can still be optimized considerably with SIMD instructions on most modern CPUs. The benefit of SATD is that it more accurately predicts quality. As such, it is often used as a way to drive and estimate distortion explicitly.

B. Motion intensity

The mean absolute difference (MAD) is the most popular block matching criterion. Corresponding pixels from each block are compared and their differences summed, as described by this equation

$$MAD = \frac{1}{mn} \sum_{p=1}^m \sum_{q=1}^n |A(p, q) - B(p, q)| \tag{10}$$

Mean Absolute Difference function for two blocks A and B of size $n \times m$. $A(p, q)$ is the value of the pixel in the p^{th} row and q^{th} column of block A.

The ratio of MAD of Adjacent frames can be used as described the motion intensity.

$$\beta = \frac{MAD_k}{MAD_{k-1}}. \quad (11)$$

When $c(x, y)$ and $L(x, y)$ are apply to video sequence, its operation was repeated for each frame and their mean values $\overline{c(x, y)}$ and $\overline{L(x, y)}$ can be computed.

The temporal information is based upon the motion difference frame, $\Delta g(m, n, t)$, which is composed of the differences between pixel values at the same location in space but at successive frames. $\Delta g(m, n, t)$ is defined as:

$$\Delta g(m, n, t) = |g(m, n, t) - g(m, n, t-1)|. \quad (12)$$

Where t is the sequence number of frame in video. The mean square error of $\Delta g(m, n, t)$ is σ_g . When the scene in the video sequence was switched σ_g is greater than others. Set T is the threshold. When $\sigma_g > T$, the scene in the video sequence was switched. Then the frame was marked K.

When the scene is switched, the difference of the contiguous frame brightness is great. Meanwhile the motion is fast. Then the sensitivity of the HVS to the distortion was weakened. Consequently, when $\overline{c(x, y)}$ and $\overline{L(x, y)}$ were calculated, the 3 contiguous frames before and after K frame were ignored. In fact, it is better to divide a video into clips according to the K frame, and then the video qualities of different clips are assessed respectively.

When the motion intensity is considered,

$$c(x, y) = \frac{k}{\beta} c(x, y). \quad (13)$$

$$\overline{c(x, y)} = \frac{k}{\beta} \overline{c(x, y)}. \quad (14)$$

$$L(x, y) = \frac{p}{\beta} L(x, y). \quad (15)$$

$$\overline{L(x, y)} = \frac{p}{\beta} \overline{L(x, y)}. \quad (16)$$

The k and p are weighting coefficients.

IV. VIDEO NOISE REDUCE

A. Video noise

Video noise reduction to increasing the user quality perception is still one of the most challenging topics [14][15]. Video noise reduction is the process of removing noise from a video signal. Video noise reduction techniques are conceptually very similar regardless of the video signal being processed, however a priori knowledge of the characteristics of an expected video signal can mean the implementations of these techniques vary greatly depending on the type of video signal. Video noise can be

random or white noise with no coherence or coherent noise introduced by the devices mechanism or processing algorithms. Typical video noise types are following:

Analog noise. These include radio channel artifacts, VHS artifacts and film artifacts.

High frequency interference, brightness and color channel interference and video reduplication are the typical radio channel artifacts. VHS artifacts mainly include color-specific degradation, brightness and color channel interference, chaotic line shift at the end of frame and wide horizontal noise strips. Meanwhile typical film artifacts are dust, dirt, spray, scratches, curling and fingerprints.

Digital noise. Typical artifacts are blocking, ringing and blocks damage.

Blocking is low bit rate artifacts. And ringing is low and medium bitrates artifact especially on animated cartoons. Blocks damage in case of losses in digital transmission channel or disk injury.

B. Video noise reduction methods

Spatial video noise reduction methods are commonly used when only one frame is used for noise suppression. Such methods are close to image noise reduction. Spatial noise reduction only uses the spatial correlation of single frame. So the effect of filter is insufficient. Sometimes the flash frames, or blurring edge or blurring texture could be appearing.

Temporal video noise reduction methods are applied when only temporal information is used [10][11]. Such method can be divided into:

Motion adaptive methods - some analysis of pixel motion detection is used. If there is no motion in some pixels - serious averaging with previous pixels are used. In case of motion more accurate averaging required to avoid "ghosting" artifacts.

Motion compensative methods use motion estimation to predict and consider right pixel values from correct place from previous frame. This method requires more time, but produce better results.

A great deal of effort has been made to develop novel video noise reduction methods. The spatial methods are based on the single frame quality, and the motion information inter frames is not considered. Therefore the results cannot be a good. The temporal methods have little effect on the motion complex video sequence. Therefore the noise reduction methods of video sequences must consider the motion information, in addition to using single-frame image processing methods. Therefore, it would be desirable to provide a method and for filtering video data to reduce noise and distortion.

The signal of current block is greatly relevant to the signal of matching block, and the noises of them are random. Therefore the noise can be reduced by calculating the average of these frames [3].

Noises such as Gaussian and impulse noise may exist in video capture because of insufficient luminance or defect of image sensors. Noise in video will seriously affect human visual perception and reduce compression efficiency. Therefore, noise reduction is a necessary part of video processing. Many spatial filter and temporal filter were proposed to reduce noise in video in previous works. Both of them have disadvantages such as ghost effect of spatial filtering and blurring in temporal filtering [13].

Consequently, we should propose a spatial-temporal filtering scheme that utilizes motion complexity with multi-reference frames in H.264 to remove noise. It utilizes the variation of inter mode distributions to detect noise and determine the parameters for spatial filter. In the time domain, a reference pixel is selected from multi-reference frames according to the motion complexity criterion. The proposed noise reduction methods can effectively remove Gaussian and impulse noises to improve video quality. Meanwhile, it can also boost the compression efficiency by reducing the bit rate.






Threshold of the motion complexity is express as TH. Then

$$\begin{cases} \text{if } \beta \leq TH & \text{spatial noise reduction} \\ \text{if } \beta > TH & \text{temporal noise reduction} \end{cases}$$

V. EXPERIMENTAL RESULTS

Five test sequences, which are representative of the specific spatial and temporal activity levels, were used to the experiment. These well known video clips are shown in table 1.

TABLE II. THE TEST VIDEO SEQUENCES

No	The Test Video Sequences		
	Activity	Name	Video
Clip1	Low	Akiyo	
Clip2		News	
Clip3		Foreman	
Clip4		Football	
Clip5	High	Stefan	

The Temporal and Spatial Activity of Akiyo is low and football is high. Each test video clip has 90 frames with 30 fps, and was transcoded by H.264 encoder to CIF format at bit rates 128K, 256K, 384K, 512K, 640K, 768K, 896K, 1024K and 1.2Mbps[9].

In order to calculate $g(m,n,t)$ and $\varphi(m,n,t)$ the frame will be filtered. The Sobel filter did not perform sufficient averaging to obtain robust estimates of the angular orientation of the spatial gradient energy. Therefore, the filter pairs in Fig. 1 and Fig. 2 were applied in the experiment procedure. The two filters are applied separately, one to enhance horizontal pixel differences while smoothing vertically (Fig. 1), and the other to enhance vertical pixel differences while smoothing horizontally (Fig. 2). The two filters were examined, and higher amounts of edge enhancement and noise suppression could produce better spatial distortion metrics than the Sobel filter.

In Fig. 1 and Fig. 2, $w_1=0.0696689$, $w_2=0.0958102$, $w_3=0.0769103$, $w_4=0.0426985$, $w_5=0.0173386$, $w_6=0.0052578$.

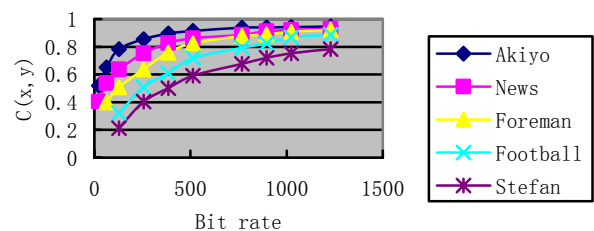
$-w_6$...	$-w_1$	0	w_1	...	w_6
\vdots	\ddots	\vdots	\vdots	\vdots	\ddots	\vdots
$-w_6$...	$-w_1$	0	w_1	...	w_6
$-w_6$...	$-w_1$	0	w_1	...	w_6
$-w_6$...	$-w_1$	0	w_1	...	w_6
\vdots	\ddots	\vdots	\vdots	\vdots	\ddots	\vdots
$-w_6$...	$-w_1$		w_1	...	w_6

Figure 1. The horizontal filter

$-w_6$...	$-w_6$	$-w_6$	$-w_6$...	$-w_6$
\vdots	\ddots	\vdots	\vdots	\vdots	\ddots	\vdots
$-w_1$...	$-w_1$	$-w_1$	$-w_1$...	$-w_1$
0	...	0	0	0	...	0
w_1	...	w_1	w_1	w_1	...	w_1
\vdots	\ddots	\vdots	\vdots	\vdots	\ddots	\vdots
w_6	...	w_6	w_6	w_6	...	w_6

Figure 2. The vertical filter.

Figure 3. The distortion of spatial information



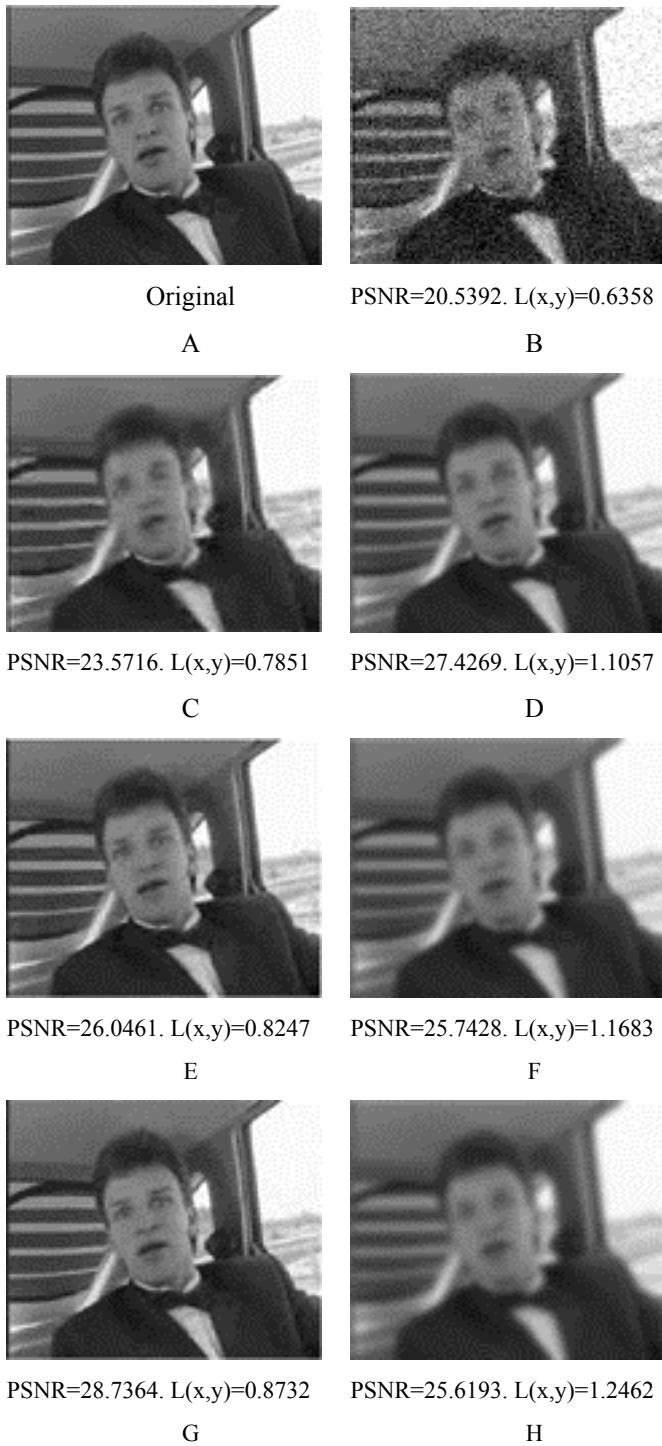


Figure 4. The type of distortion

The experiment on $c(x, y)$ was carried out with these video. The results were depicted in Fig. 3.

Comparing the curves of Fig. 3, it is deduced that higher spatial and temporal information results in higher distortions at the same bit rates. Meanwhile, the lower spatial and temporal information needs lower bit rates at the same distortion. The $C(x, y)$ can be used to quantify the distortion of the spatial information.

In order to specify the effect of $L(x, y)$ across different distortions, a serial of images is used to the

experiment procedure. The original image is the 51th frame in Carphone video. The results were shown in Fig. 4. In Fig. 4, the images C,E,G are blocking distortion, and image B is blocking and Gaussian noise distortion, D, F, H are blurring distortion. These distortions are made by encoding and adding noise to the original images.

Comparing the $L(x, y)$ values of Fig. 4, it is deduced that blocking-distortion results in small $L(x, y)$ value, and blurring distortion results in greater $L(x, y)$ value. Thus the distortion can be distinguished between blocking and blurring.

TABLE II. LOW MOTION INTENSITY TESTING VIDEO SEQUENCE

No	The Test Video Sequences		
	original	Noise	Filtered
120			
125			
130			
135			
140			
















Two test video sequences, which are representative of the specific spatial and temporal activity levels, were used to the experiment. These well known video clips are shown in Fig.1 and Fig.2.

The temporal and spatial motion complexity of sequence News is low and the football sequence is high. Each test video clip has 90 frames with 30 fps, and was transcoded by H.264 encoder to CIF format at bit rates 1024K.

From the Fig.1 and Fig.2, we can find that the method of noise reduction is good for both the low motion complexity and high motion complexity. The proposed noise reduction methods can effectively remove Gaussian and impulse noises to improve video quality. Meanwhile, it can also boost the compression efficiency by reducing the bit rate.

Filtering can also tax system throughput, since increased computational complexity often results from filtering schemes. Furthermore, the movement of objects within frames, as defined by groups of pixels, complicates the noise reduction process by adding additional complexity. In addition to improvements made to FIR spatial filtering, the present invention improves on previous filtering techniques by using Infinite Impulse Response temporal filtering to reduce noise while maintaining edge definition.

TABLE III. HIGH MOTION INTENSITY TESTING VIDEO SEQUENCE

No	The Test Video Sequences		
	<i>original</i>	<i>Noise</i>	<i>Filtered</i>
11			
13			
15			
17			
19			

VI. CONCLUSIONS

In this article, we proposed image quality measures $c(x, y)$ and $L(x, y)$. They could quantify the spatial distortion and the type of the distortion. Thus, perceived spatial information can be measured approximately. Because perception of human has been taken into account, the metrics correlate well with the subjective quality measures. Experimental results show the validity of our approach and our perceptual quality metric performs better than the existing methods.

ACKNOWLEDGMENT

This work was supported by the Zhejiang Province Natural Science Foundation of China under Grant No. X106870.

REFERENCES

[1] K. Seshadrinathan, and A.C. Bovik, "A structural similarity metric for video based on motion models," IEEE

International Conference on Acoustics, Speech, and Signal Processing, Honolulu, Hawaii, 2007, pp. 1-869-72.
 [2] A.B. Watson, J. Hu, and J.F. McGowan III, "Digital video quality metric based on human vision," J. Electron. Imaging, vol. 10, no. 1, Jan. 2001, pp. 20-29,
 [3] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in Proc. IEEE Int. Conf. Image Processing, 1994, pp. 982-986.
 [4] Z. Wang and A. C. Bovik, "A universal image quality index," IEEE Signal Processing Letters, vol. 9, no. 3, Mar. 2002, pp. 81-84.
 [5] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," Signal Processing: Image Communication, vol. 19, no. 2, Feb. 2004, pp. 121-132.
 [6] K. Seshadrinathan and A. C. Bovik, "Statistical video models and their application to quality assessment," in Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, AZ, Jan.2006, pp 23-26.
 [7] A. A. Webster, C. T. Jones, M.H. Pinson, S. D. Voran, S. Wolf, "An objective video quality assessment system based on human perception", in SPIE Human Vision, Visual Processing and digital display IV, vol. 1913, San Jose, CA, Feb. 1993 , pp. 15-26.
 [8] Z. Wang, R. Hamid, and A.C. Bovik, The handbook of video database : design and applications, CRC Press , New York , 2003.
 [9] <http://live.ece.utexas.edu/research/quality/>
 [10] S.-W. Lee, V. Maik, J. Jang, J. Shin, and J. Paik. "Noise adaptive spatio-temporal filter for real-time noise removal in low light level images," IEEE Tran. Consumer Electronics, vol. 51, no. 4, May 2005, pp. 648-653.
 [11] K. Miyata and A. Taguchi, "Spatio-temporal separable data-dependent weighted average filtering for restoration of the image sequences," IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4,2002, pp. IV-3696-IV-3699.
 [12] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," IEEE Trans. Circuits System Video Technology, vol. 13, July 2003, pp. 560-576.
 [13] M. Boyce, "Noise reduction of image sequences using adaptive motion compensated frame averaging," IEEE Int. Conf. Acoust. , Speech, Signal Process. , 1992 , pp. 461-464.
 [14] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in Proc. IEEE Int. Conf. Image Processing, 1994, pp. 982-986.
 [15] Z. Wang and A. C. Bovik, "A universal image quality index," IEEE Signal Processing Letters, vol. 9, no. 3, Mar. 2002, pp.81-84
 [16] C.Yang, Y.Liu and J.Yu, "Objective Quality Metric Based on Spatial-temporal Distortion," IEEE International Conference on MultiMedia and Information Technology, 2008,pp.813-816
 [17] C.Yang, L.Zhao and Z.Liao, "Objective Quality Metric Based on Perception for video," IEEE International Conference on Computer Engineering and Technology,2009,pp.20-24.
 [18] C.Yang, Y.Liu and J.Yu, "Research on Video Transmission Based on Motion Intensity," IEEE International Conference on Information Technology for Manufacturing Systems,2010,pp.81-85



Chunting Yang was born in Qiqihar City, P. R. China, on January 16, 1964. Yang received a BS in 1986 from Nanjing University of Aeronautics and Astronautics. Then he received his MS in 1991 and PhD in 1996 from the Southeast University and Zhejiang University. He is an associate professor of computer science and engineering at the Zhejiang

University of Science and Technology. His research interests are in computer image processing, computer vision and speech signal processing.

Dr. Yang is the member of a council of Zhejiang Computer Society and the member of a council of Zhejiang Electronics Society.



Yang Liu was born in Wuhan, P. R. China, on June 4, 1978. Liu received her BS in 2000 and MS in 2003 from Huazhong Normal University.

She is a full-time lecturer of computer science and engineering at the Zhejiang University of Science and Technology. Her research interests are in computer vision, virtual reality.



Jing Yu received her PhD in Mechanical Engineering in 1998 from Zhejiang University (Hangzhou, China) and worked as an Associate Professor in the Institute of Vibration Engineering Research at Nanjing University of Aeronautics and Astronautics (Nanjing, China) for the period of May, 1998 to March, 2000. Between October

2000 and April 2002, she worked as a Research Associate in Lakehead University, Canada. At present, she is engaged in research in University of Toronto, Canada. Her areas of research interest include random vibration control and signal processing..

Tele-Immersive Interaction with Intelligent Virtual Agents Based on Real-Time 3D Modeling

Shujun Zhang

¹College of Information Science & Technology, Qingdao University of Science & Technology, Qingdao 266061, China

²State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China

Email: lindazsj@163.com

Wan Ching Ho

Adaptive Systems Research Group, School of Computer Science,
University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB, UK

Email: w.c.ho@herts.ac.uk

Abstract—To enable intelligent agents interacting smoothly with human users, researchers have been deploying novel interaction modalities (e.g. non-verbal cue, vision and touch) in addition to agents' conversational skills. Models of multi-modality interaction can enhance agents' real-time perception, cognition and reaction towards the user. In this paper we report a novel tele-immersive interaction system developed using real-time 3D modelling techniques. In such system user's full body is reconstructed using multi-view cameras and CUDA based visual hull reconstruction algorithm. User's mesh model is then loaded into a virtual environment for interacting with an autonomous agent. Technical details and initial results of the system are illustrated in this paper. Following that a novel interaction scenario is proposed which links the virtual agent with a remote physical robot who takes the role of mediating interactions between two geographically separated users. Finally we discuss in depth the implications of such human-agent interaction and possible future improvements and directions.

Index Terms—intelligent virtual agent, tele-immersion, interaction, real-time modeling

I. INTRODUCTION

Intelligent agents are interactive virtual characters or physical robots that can exhibit human-like qualities and communicate with human users and other agents through multi-modalities such as vision, gesture and speech. These agents, which are either inhabiting simulated virtual environment (VE) or the real world, require an integrated perspective of embodiment in communicating with users [1]. Following this direction researchers in human-agent interaction (HAI) have explored a wide range of communicative functions including language and bodily actions to create natural and advanced interfaces to which users can easily adapt. Studies on the recognition and comprehension of non-verbal signals are particularly challenging – one of the reasons is that user's body movements (e.g. hand gestures and head movements) are difficult to be translated precisely to

represent consistent communicative signals understood by agents [2,3].

As non-verbal signals convey a multitude of information which can be exploited for HAI, in recent years they are considered as an extremely rich source to create intelligent agent interface. One direction, which is based on literatures on human communicative behavior [4,5], is to create multimodal corpora. Since a corpus is “a collection of video recordings of human (or human-agent) communicative behavior that is annotated or coded with different types of information ([6, page 3])”, it helps humans and in particular intelligent agents to identify correctly to a certain extent the communicative signals from each other. Another direction is to use advanced sensory technology to track user's bodily actions including user's gestures, head movements, eye gazes and even emotions (see an example in [7]). However, to interact with those systems the user whose actions or expressions need to be recognized are required to wear or carry a device (e.g. track-able marker or controller [8]) and each device senses only the actions performed by a particular body part. Although recently Microsoft® Xbox 360 has provided a tracking device Kinect¹ for gaming environment, which is claimed to be able to recognize full human body movements, the official availability of the resources for this device used in academic research is still unknown.

To further investigate the research direction of “markerless” HAI, in this paper we propose a new communicative interface system which offers user an enhanced tele-immersive experience while interacting with autonomous agents. The system requires no sensors attached to the user and it creates a simulated 3D model of the user's full-body and motions in real-time. Meanwhile the user's 3D model gets exported to a VE where it can interact with intelligent agents embodied in the environment through natural gestures and movements.

The paper is structured as follows: In Section 2 research work related to simulations of real-time 3D

¹ <http://www.xbox.com/en-US/kinect>

model in mixed reality (MR) and tele-immersive interactions are described. Section 3 and 4 first provide the main concept and principles of the research work, and then illustrate our system's initial implementation and some preliminary results. Finally in the Discussions section we summarize the technical contributions of the paper and provide a list of limitations which indicate existing questions that still need to be answered.

II. RELATED WORK

A. Image-based 3D Modeling

In the research field of MR, image-based 3D modeling has been an important direction fusing the "real" into the "virtual". In order to achieve real-time 3D reconstruction, research groups at various universities around the world have studied modeling techniques and developed different image-based modeling systems. The representatives include GrImage at INRIA [9], tele-immersive system in the DISCOVERY lab at University of Ottawa [10], Virtualized Reality system at Carnegie Mellon University [11,12] and Keck Laboratory at the University of Maryland [13].

GrImage at INRIA uses EPVH (Exact Polyhedral Visual Hull) method [14,15] and cluster parallelization techniques for real-time reconstruction which requires a large cluster of computers [16]. The DISCOVERY lab at University of Ottawa developed a tele-immersive system using also EPVH but resulted in lower modeling efficiency [10]. Virtualized Reality system at CMU first used multi-baseline stereo techniques to extract models of dynamic scenes [11,12] and then they explored volume-carving methods with epipolar geometry to generate an image of the desired view [17,18]. The major issue of their techniques is that it requires a multitude of off-line processing. The system built by Keck Laboratory at University of Maryland is highly sophisticated with three-room scale and the modeling effect is low in accuracy [13].

In contrast with the above modeling techniques and systems, the State Key Laboratory of Virtual Reality Technology and Systems at Beihang University in China proposed a CUDA²-based visual hull modeling technique and created a MR system named "DreamWorld" for real-time 3D modeling and interaction [19]. DreamWorld makes use of the parallel computation power of GPU to effectively improve the modeling speed, so that the system can support real-time reconstruction of the full-body of multiple users. Meanwhile, it can produce an explicit mesh model which is easy to be exported for other applications. We used DreamWorld system for 3D modeling in this paper.

B. Tele-Immersion

With the development of multi-media and virtual reality technology, tele-immersion systems are attracting more and more attention from both academia research

and industrial application development. Generally, there're two types of tele-immersion systems. The first type combines direct local modeling and remote transmission, and the second type is based on feature extraction and avatar substitutes. Although the latter can reduce the scale of the data transmission, the trade-off is the inferior immersive effect. Two known efforts of the former type are tele-immersion system at UC Berkeley [20] and SVTE (Shared Virtual Table Environment) system [21] at Heinrich-Hertz Academy. UC Berkeley uses stereo-based modeling method to get a point cloud model of users' body [20]. It can produce a good visual effect due to the large multitude of the point cloud, however, it can't support virtual-reality interaction due to the fact that discrete points cannot model smooth surface. Heinrich-Hertz Academy proposed the idea of SVTE and developed a virtual meeting room using quasi-3D modeling technique [21]. In SVTE users' front images are loaded into a VE to produce a feeling of face-to-face communication. This method is similar to video conference. As for the second type, Twente University used virtual avatar to represent the local user for tele-immersion after extracting and quantifying the features of the local users' gestures and behaviors [22,23]. Compared to the real-time modeling system, this method loses part of fidelity in creating the VE. All the above research aims at setting up peer-to-peer direct communication among users who are geographically separated. To the best of our knowledge, little research has been carried out using tele-immersive systems for remote social interactions between users and intelligent agents.

C. HAI in Mixed Reality Environments

In MR systems, HAI can be diverse. Through different kinds of interaction modality, the agent can learn a rich set of behaviors and thus improve its adaptability to the user and the environment. Duffy and colleagues created NEXUS [24] - a framework that supports fusion of physical and VEs, in which people can interact with deliberative agents in their own space. He utilised JARToolkit [25] to support video input and the sensing of events in the physical world. The principle of this augmented reality system is rendering virtual characters into the place where the real markers mapped. Researchers in [26,27] introduced the notion of "MR agent" - an agent consisting of a physical robotic body and a virtual avatar displayed upon it. They also developed an interface using a head-mounted display to interact with such systems. A framework for explicit deliberative control of socially and physically situated agents in virtual, real and MR environments has been described in [28]. Similarly, research shown in [29] has developed a platform that enables a virtual character to display blended facial expressions with emotions as real-time continuous reactions to users' gesture input. Finally, a virtual reality system was set up in [30], allowing a human player to be immersed in a CAVE and interact with the virtual multi-agent team in RoboCup. The system could also support different users at different geographic locations to participate in the game in real time. In this system users had to wear three trackers to

² CUDA is NVIDIA's parallel computing architecture:
<http://www.nvidia.com/>

allow their motions to be transferred separately in head,

III. CONCEPT AND PRINCIPLE OF THE PROTOTYPE

To focus on the remote communication issues for multiple users and agents, and to expand the flexibility of motion and gesture recognition in HAI, here we propose a tele-immersive interaction prototype with autonomous agents based on real-time 3D modelling. Our prototype has the following characteristics:

(1) Using MR object modelling, especially in real time: The system can reconstruct the user's body into mesh model and embed it into a VE. This technique creates a new 3D interaction modality.

(2) Tele-immersive interaction: Users can see themselves in the VE and "touch" the agent as if they coexist in the same location. Both the vivid model of the full-body of users (including recognisable gesture and/or movement) and the instant feedback from the agent contribute to the immersive effect.

(3) Markerless and free input mode: In the previous research work of HAI, users usually need to wear certain equipments (e.g. data gloves), and that creates an explicit limitation of free movements as well as naturalness. We use camera-based passive data acquisition and computer vision based processing method to offer a markerless and free input mode for the user.

In addition to the fact that the HAI prototype in this paper is focused on the interaction between the user and the autonomous agent in the VE, the framework can also be extended through network to support the communication with other remote intelligent agents and their own users. Therefore, three kinds of interaction chain are formed: the local user (the full or part of body is modelled by DreamWorld and loaded into Webots [31]) with the virtual agent in Webots, the remote user with another agent and the two users. The concept of system scenario is shown in Fig.1.

In Fig.1, the left two pictures show the modeling platform which is called DreamWorld. The user's body can be reconstructed using DreamWorld in two different levels. The upper one is smaller in scale and it can precisely model part of the user's body such as a hand; the lower one supports full-body reconstruction. The models output from either DreamWorld will be loaded into Webots to build our MR system, as shown in the middle graph. Virtual robot (AIBO ERS-7 for example)

wand and foot. This may limit users' movement. in Webots' environment will react to the input model (e.g. a reconstructed 3D human hand). Since the model is created in real time and loaded dynamically, the VE is changing continuously in the eye of the robot. Thus the user can produce corresponding behaviors and through Internet, his/her status will be transferred to the remote agent (e.g. a physical robot). Finally, at the right figure, remote user B with the physical robot gets the information of user A and thus forms a new type of interaction modality (a prior research using only physical robots can be found in [32]). Similarly, the physical robot's internal states and behaviors executed can also be sent to the remote virtual agent in Webots, and user A may react differently when he sees virtual AIBO's dynamic response which can be synchronized with the remote robot.

The main strengths of the interaction scenario are:

(1) The sensor-free interaction mode is more natural and comfortable to the users.

(2) It supplies a new way for users to interact with a remotely located robot. They can "reach" the robot any time at any remote place with virtual feedback. Similarly, users can also remotely teach skills to or tele-control their robots using the new MR system.

A. Real-Time 3D Modeling Using DreamWorld

DreamWorld is a platform that consists of hardware and software supporting real-time 3D modeling, developed by the State Key Laboratory of Virtual Reality Technology and Systems at Beihang University. There are two configurations according to different scale of acquisition space and number of cameras.

(1) DreamWorld-I is of 1m×1m×1m scale, made up of a plank cube with five cameras, five PCs, one graphic workstation and three displays. It is suitable for modelling and interaction of static objects and part of users' body movement such as hand or arm.

(2) DreamWorld-II is of 3m×3m×3m scale with 24 cameras and it supports multi-users' full body data acquisition. The cameras can be selected according to required viewpoints. In this platform, multiple users can behave freely and naturally, interact with each other and be reconstructed in real time.

The two settings are shown in Fig. 2.

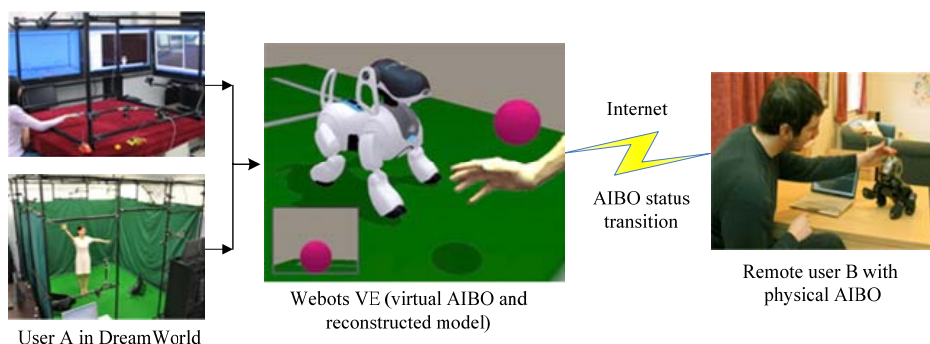


Figure 1. System scenario

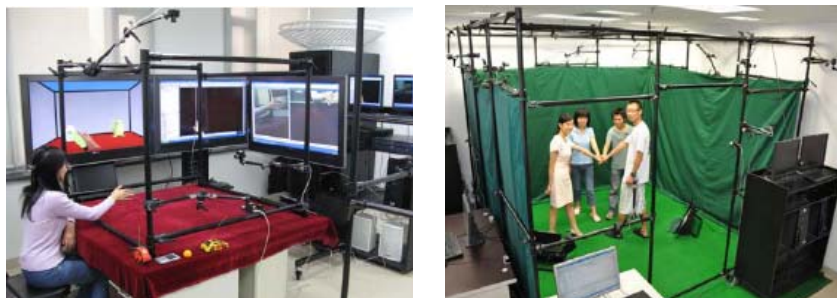


Figure 2. DreamWorld hardware

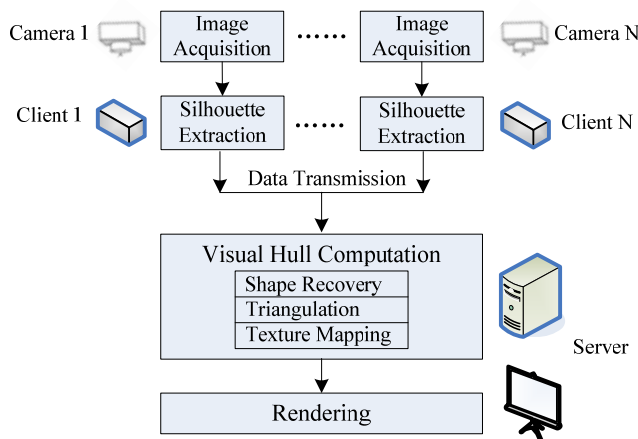


Figure 3. System architecture of DreamWorld

Through network communication, the two sets of software operating on the two platforms in Fig. 2 can be connected into an integrated system, which supports physically separated multiple users sharing data and interacting with the same virtual environment. Users can be immersed into a dreamed world according to the application type when wearing helmet display.

The system architecture (Fig. 3) is built in Client/Server mode. Each client PC is connected with a camera to capture images of the object from one viewpoint and all the clients complete silhouette extraction from the images with synchronization. The server is in charge of visual hull reconstruction using the received multi-view silhouette images got from all clients.

In Fig. 3, there are five basic steps to accomplish the complete reconstruction task: image acquisition, silhouette extraction, data transmission, visual hull computation and rendering. In visual hull computation, there are three sub-procedures which are all implemented in CUDA parallelization: shape recovery, triangulation and texture mapping [19].

Visual hull modeling is one of the image-based 3D reconstruction methods. The definition of “visual hull” is the largest intersection of viewing cones and it can be computed using 2D silhouette information from each image. The procedure of getting the visual hull of a target object is this: dividing the initial modeling space into different cubes (called “voxels”) and projecting each cube into every image to test whether it belongs to the object or not. If the voxel’s projection lies completely inside all the silhouettes, then the voxel is marked as Black; if the

voxel’s projection lies completely outside any one of the silhouettes, the voxel is marked as White; otherwise (which means the projection is partly inside the silhouettes), the voxel is marked as Gray. Through deleting the voxels which don’t belong to the object from the initial cube, voxels belonging to the object can be got. And the Gray cubes can be further divided into small cubes until the required precision. Since we use octree structure to complete this procedure, the MAX_DEPTH of the octree decides the modeling precision and also influence the efficiency of the system. The core computation step here is to carry out the intersection judgment of each voxel’s projection with all the silhouettes.

After all the voxels are tested, those voxels marked as Black constitute the visual hull of the object and particularly, the Gray ones form the boundary of the visual hull. Since only the voxels on the margin contribute to the shape of the object, the Black ones are deleted which are inside the object so as to reduce the amount of voxels that need to be processed in triangulation phase. Triangulation is executed using marching cubes algorithm to get a smooth mesh surface of the object. The principle of marching cubes is finding the exact intersection points of the viewing cones and every silhouette, then constructing iso-surface over the intersection points. All the iso-surfaces form the final mesh of the target object.

And finally, the mesh structure needs to be textured using the source image information with exact coordinate mapping. Texture is a very important factor for the visual effect of the reconstruction result. We used view-

dependent texture mapping method and shader in GPU for acceleration.

B. Intelligent VE (Webots Software)

While the focus of this research is not the development of our own graphic engine, using a stable simulated VE, such as Webots [31], is essential for us to have various virtual robot models which are virtually identical to the corresponding physical robots. Therefore the data received from the virtual sensors can be transferred between virtual and physical robotic platforms.

Through utilising the Supervisor Node from Webots' modules, we can load the reconstructed model from DreamWorld dynamically into Webots. The model which reflects the reality and the other virtual objects together constitute a VE for the virtual agents in Webots. As physics simulation is supported in the Webots' environment, collision and friction can be detected from interaction between the reconstructed human body (or hand) model and virtual agents (and objects) created in Webots directly. Thus intelligent virtual agents in Webots can react to the user's movement and gesture (captured in DreamWorld) in various ways as designed in the agent architecture.

C. Remote Interaction Scenario

Nowadays, more and more people live away from their families and friends either because of their work commitments or due to other personal reasons. The current up to date means of communication include appliances such as phones or other conferencing devices such as an online computer equipped with a video camera. However, regardless of the fact that voice/video communication exists, it still lacks some significant ways and enjoyable factors allowing people to be more engaged while interacting with each other. One of the authors' previous research used physical robots acting as social mediator to enrich such remote interaction with the inclusion of the tactile sensation (touch) that could be transferred over distance [32].

Human's full body movements may provide more interactive and enjoyable communications if the means of interaction are designed appropriately and acceptable from the users. Therefore, the proposed immersive virtual system can have positive implications for user-computer and user-user interactions, especially in the context of remote communication or communication within a gaming environment. Authors believe that the realization of richness derived from body movements will further increase the engaging experiences of users.

Towards realising and verifying this scenario setting in the near future, we aim to adapt the system prototype for remote interaction between geographically separated users. As proposed in the previous subsections, we use DreamWorld to model one or both of the users, thus allowing him/her to interact with the VE which increases the dynamics of interaction, such as moving their body around or manipulating virtual objects. For example, as illustrated in Fig. 1, the virtual robot which is synchronised with the remote physical robot can produce feedback to both users' action in the VE. At the same time, the remote user can experience the feedback provided by the physical robot during the interaction. Through this design of the new remote communication, users' interaction and telepresence can be enriched and mediated by the virtual and physical robots [32,33].

IV. RESULTS FROM PRELIMINARY IMPLEMENTATION

As preliminary implementation was carried out for creating the prototype of our MR system, we have also obtained some test results from our system. Here we describe our first set of results generated from the system.

A. Testing Environment

The MR system ran both on DreamWorld and Webots at the same time for data acquisition, model reconstruction and HAI. The two platforms were connected by TCP/IP protocol using windows socket programming. We used DreamWorld-II with 6 cameras for full body reconstruction. In order to support dynamically changeable scenery management, we chose Webots 6.2.2 Professional. Hardware and software configurations are shown in Table.1.

B. Preliminary Results

First, we illustrate the modelling result of our MR system. The user stood at the center of DreamWorld-II (Fig. 4(a-b)) and his full body was reconstructed into a 3D model in real time (Fig. 4(c-d) are pure mesh models and Fig. 4(e-f) are models after texture mapping).

The model was then loaded into Webots for real-time interaction with the virtual robot, which can detect the collision autonomously and react to the user's model, for example, changing the original path towards the destination. The process of the Webots virtual environment's change is shown in Fig. 5, where the blue colour cylinder-shaped object means the vehicle robot and the user's model lies in the white edge formed cuboid.

TABLE I. HARDWARE AND SOFTWARE CONFIGURATION

Item	Description
Client	Intel Core 2 Duo E6400 2.13 GHz, 2GB DDR2-667 memory, WINXP PRO 32, IDE: VS2005 (C++), OpenCV 1.0
Server	Intel Core 2 Duo E6400 2.13 GHz, 2GB DDR2-667 memory, WINXP PRO 32, NVIDIA GeForce 9800 GTX+, 512M. IDE: VS2005(C++), OpenGL1.0, Glut 3.7.0
Cameras	FL2G-13S2C-C with 4mm lens of FV0420, resolution 640x512, 15FPS
Network	Gigabit Ethernet
Simulation Environment	Webots 6.2.2 Professional

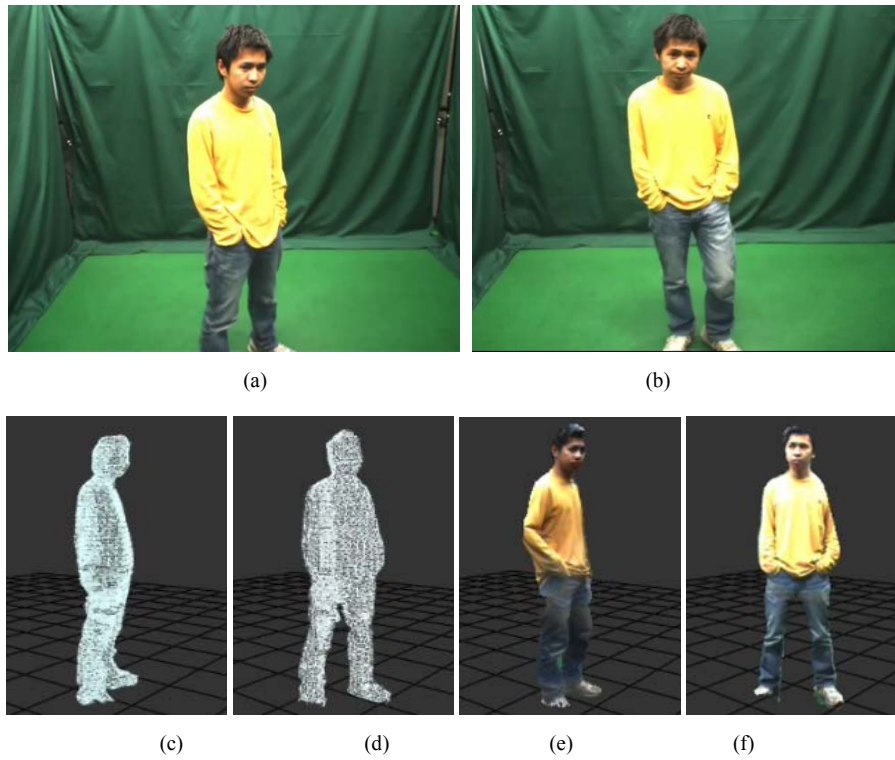


Figure 4. Modelling result

(a)The original image; (b)The user can freely change his gestures; (c) and (d) Reconstructed models from different views in DreamWorld; (e) and (f) Models with texture mapping.

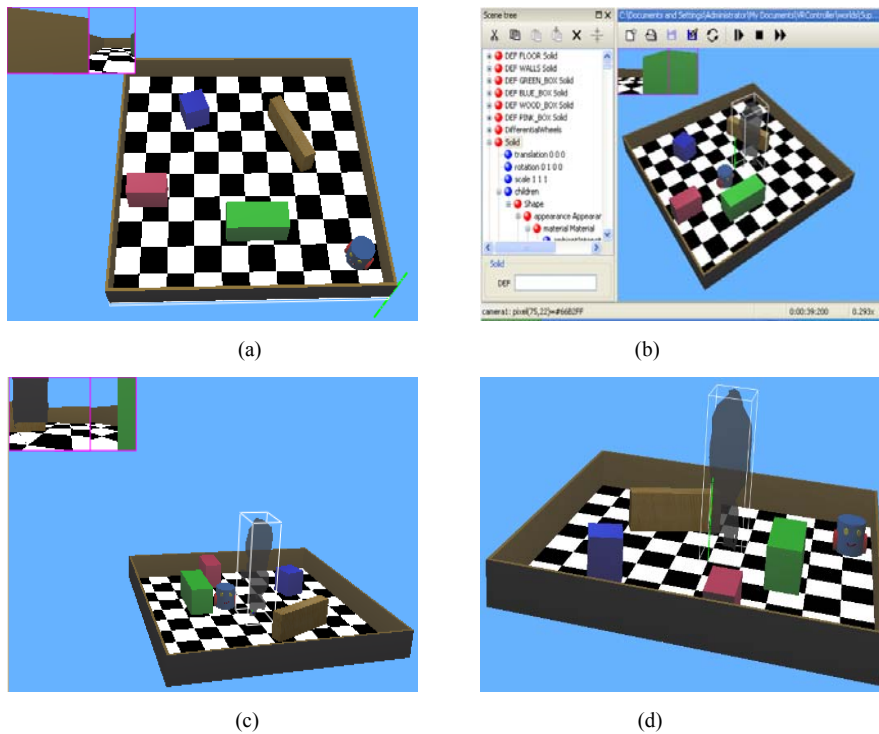


Figure 5. Preliminary test result

(a) The initial Webots environment and the left-up part is the field of vision of the robot; (b) The VE when the model is loaded in from DreamWorld and the left part is the scene tree of webots; (c) The robot finds the model of the user on its forward way; (d) The robot changes the path accordingly to avoid collision with the user.

TABLE II. PERFORMANCE OF THE SYSTEM

Item	MAX_DEPTH=5	MAX_DEPTH=6
Acquisition speed	15fps	15fps
Modeling time	67ms	257ms
Loading time	0.5s	1.4s
Number of triangles	3064	13249

From Fig. 4(e-f), the body of the user can be well modelled with satisfactory texture mapping (TM). However, in Fig. 5(b-d), the model was loaded into Webots and the texture information of the model was lost. The reason is that DreamWorld uses GPU based texture mapping method (e.g. pixel shader) which is difficult to be transferred into Webots at the current stage. The core work during the first research session is to connect DreamWorld and Webots, test the collision-avoidance function and realize the remote motion transfer function. Details and our future work addressing this issue will be described in Section 5.

With the hardware and software specifications provided above, we test the system performance including acquisition speed, modelling time, loading time (from DreamWorld to Webots) and number of triangles of the reconstructed model. The average data from ten experiments is shown in Table. 2.

As Table 2 shows, the depth of the octree had great influence on the modelling time and follow-up performance indexes. The modelling speed on average was 9 FPS. In the case of having more details of the user, more graphical depth would be needed, which would make the delay (from the input to Webots to its output) over 1 second, or even to 2 seconds. On the other hand, faster loading speed is essential because the user located at the remote location (based on the scenario we proposed in Subsection III.B) could observe his/her robot's synchronized gesture/motion and react timely. Therefore, the delay reduction will be one of the main directions in our further research.

V. DISCUSSIONS

Through the initial testing and results, we obtained the following findings:

(1) It is feasible to load a real-time reconstructed 3D model into a simulated VE for human-robot interaction. Through DreamWorld, full-body reconstruction and interaction is realized. Real-time 3D modeling can capture the changes of human users' gesture and motion in time and from different viewpoints. Therefore users can interact with the robot using a natural modality.

(2) The VE provided by Webots is able to import 3D models dynamically. In addition to the design of robots' control and behavior, it also provides an ideal platform for HAI research with intelligent virtual agents.

(3) Compared to direct interaction between a user and a physical robot, indirect interaction (such as "remote touch" or tele-immersive teaching) illustrated in this paper is a new interaction modality which opens the

research field of HAI.

In summary, tele-immersive interaction with intelligent virtual agents and/or physical robots based on real-time 3D modelling can have many potential applications:

(1) It can be used to study different HAI modalities for socially mediated remote communication between users (details of a recent study in this direction can be found in [32]).

(2) It can be used for teaching agents social skills (e.g. proxemics [34]) and gestures using remote imitation learning [35]. Such teaching methods to virtual agents or physical robots tend to be flexible and instant for the "teacher" (user) and thus our MR system can be used.

Despite the above findings and potential applications, in this research we have also encountered several limitations which need to be addressed in our future work.

First of all, although Webots allows model importation, it limits the frame rate and the size of the input file. Models loaded into Webots cannot be too large and in high frame rate because the scene and object construction in Webots is based on VRML 2.0 format. In our initial test, if the model of the human body is reconstructed in over 10 FPS or it includes over 9000 triangles, it cannot be loaded into Webots – the loading procedure can be broken down. Meanwhile, the Supervisor Node of Webots can only be modified to some extent: Collision detection of the robot at the moment can only use bounding box method. It means that modelling some specific gestures performed by the user becomes difficult and the virtual agent cannot recognise them. Therefore here we have experienced certain limitations of Webots – however, it is still a relatively stable platform that we can find in the research area.

Secondly, the visual effect of the loaded model in Webots is not satisfactory. This is due to two reasons. The first one is hardware limitation: As to DreamWorld II, the actual space to install all the cameras is only 3m×3m×3m. Although we use wide-angle (4mm) lens for 3D modelling, the effective acquisition space is still very limited since the reconstruction is based on the intersection of all the cameras' visible range. Therefore, the method cannot apply to movements involved long distance in the physical space (e.g. walking from one side of the room to another). Second, the model loses texture information after being loaded into Webots since DreamWorld uses GPU based texture mapping method which is difficult to be transferred into Webots. Fig.6 shows the texture mapping method during reconstruction

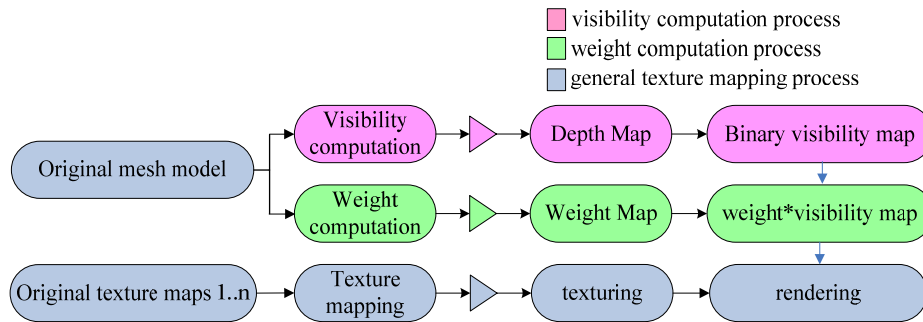


Figure 6. Texture mapping method that DreamWorld used

Fig.6 describes the three steps of the texture mapping (TM) method that DreamWorld uses: 1) Visibility computation process (first row) and 2) weight hybrid process: through weight computation a weight map is formed and the weight*visibility map is created (second row); and finally 3) pixel shader is used for rendering in texture mapping (third row). Different from general TM process - this method takes both visibility computation and weight computation into account. The view-dependent method will result in vivid texture mapping and the whole procedure is conducted in GPU, which is impossible for Webots rendering at the current version.

To solve this particular TM problem, in future research we will apply view-independent texture mapping with no shader use and add this information into the mesh model, although this could produce another problem of texture trembling to some extent.

In addition to the above refinements of our MR system, we also intend to explore different types of feedback appropriate to it. Currently, the user is limited to get other feedback in addition to the visual one from the agent's reaction. Visual feedback is one of the most important aspects in HRI, but it's clearly not enough. One feasible improvement is the use of head mounted display to replace the current large screen output, and also haptic devices to produce feedback information for enhancing the immersive level of the HAI.

VI. CONCLUSION AND FUTURE WORK

In this paper we explored the combination of real-time 3D modelling and HAI and proposed a prototype of mixed reality system for tele-immersion. Image-based real-time 3D modeling techniques offer essential ways to support tele-immersion and natural HAI. The mesh model reconstructed in DreamWorld II platform is loaded into a dynamic virtual environment, together with an autonomous virtual agent, created using the Webots software. Then the agent's behavior can potentially be copied to or influence the corresponding remote physical robot, as proposed in the user interaction scenario. Similarly, states and behaviors of the remote physical robot can be transferred back to the virtual agent and then the local user can see immediate feedback from the remote user accordingly. Therefore, tele-immersive user-user and user-agent remote interactions are achievable through the proposed MR system. As models of multi-

modality interaction can enhance agents' real-time perception, cognition and reaction towards the user, in the near future we hope that our research can be applied to the areas of remote social communication mediated by agents and distant robot teaching.

ACKNOWLEDGEMENT

We would like to thank the UK Royal Society for the 1 Year International Joint Project (UK and China) grant. We also thank Prof. Wei Wu at Beihang University for supplying the necessary resources and Prof. Kerstin Dautenhahn at University of Hertfordshire for her comments and supports. This paper is partially supported by the National Natural Science Foundation of China (Grant No. 60903064 & 61040047), the National Grand 973 Program of China (Grant No. 2009CB320805), and the European Commission fund (EU FP7 ICT-215554 project LIREC)

REFERENCES

- [1] I. Wachsmuth, G. Knoblich, (Eds.): *Modelling Communication with Robots and Virtual Humans, Second ZiF Research Group International Workshop on Embodied Communication in Humans and Machines*, Bielefeld, Germany, April 5-8 (2006)
- [2] M.L. Knapp, J.A. Hall, *Nonverbal Communication in Human Interaction*. (5th ed.) Wadsworth: Thomas Learning. ISBN 0-15-506372-3 (2007)
- [3] T. Sowa, *The Recognition and Comprehension of Hand Gestures - A Review and Research Agenda*. In *Modeling Communication with Robots and Virtual Humans*. Springer Berlin/Heidelberg. pp. 38-56 (2008)
- [4] D. McNeill, *Hand and Mind-What Gestures Reveal about Thought*. The University of Chicago Press, Chicago, London (1992)
- [5] A. Kendon, *Gesture-Visible Action as Utterance*. Cambridge University Press, Cambridge (2004)
- [6] M. Rehm, E. André, From Annotated Multimodal Corpora to Simulated Human-Like Behaviors. in *Modelling Communication with Robots and Virtual Humans*, I. Wachsmuth and G. Knoblich, Springer, pp.1-17 (2008)
- [7] M. Dunne, B.M. Namee, J. Kelleher, Intelligent Virtual Agent: Creating a Multi-Modal 3D Avatar Interface. In *Proceedings of the 9th Annual Information Technology & Telecommunication Conference (IT&T '09)*. (2009)
- [8] R Y. Wang, J. Popovic, Real-Time Hand-Tracking with a Color Glove. *ACM Transaction on Graphics (SIGGRAPH)*, New York, USA, 28(3):1-8 (2009)

- [9] J. Allard, C. Menier, B. Raffin, et al. Grimage: Markerless 3D Interactions, In *ACM SIGGRAPH'07, International Conference on Computer Graphics and Interactive Techniques, emerging technologies*, article No. 9 (2007)
- [10] C. Leong, Y. Xing, N.D. Georganas, Tele-Immersive Systems. *IEEE International Workshop on Haptic Audio Visual Environments and Their Applications*, Ottawa: Canada, (2008)
- [11] P. Narayanan, P. Rander, T. Kanade, Constructing Virtual Worlds Using Dense Stereo. *International Conference of Computer Vision (ICCV98)*, pp. 3-10 (1998)
- [12] T. Kanade, P.W. Rander, P.J. Narayanan, Virtualized Reality: Constructing Virtual Worlds from Real Scenes. *IEEE Transactions on Multimedia*, Vol.4, Issue 1, pp.34-47 (1997)
- [13] E. Borovikov, L. Davis, A Distributed System for Real-time Volume Reconstruction. Proceedings of the Fifth *IEEE International Workshop on Computer Architectures for Machine Perception*, Padova: Italy, pp. 183-189 (2000)
- [14] J.-S. Franco, E. Boyer, Exact polyhedral visual hulls. Proceedings of the *Fourteenth British Machine Vision Conference (BMVC)*. Norwich, UK: BMVA Press, pp. 329-338 (2003)
- [15] J.-S. Franco, E. Boyer, Efficient Polyhedral Modeling from Silhouettes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Volume 31, Issue 3, 414-427 (2009)
- [16] J.-S. Franco, C. M'E Nier, E. Boyer, et al: A Distributed Approach for Real Time 3d Modeling. Proceedings of the *IEEE Workshop on Real Time 3D Sensors and Their Use* (2004)
- [17] M. Kimura, H. Saito, T. Kanade, 3D Voxel Construction Based on Epipolar Geometry. *International Conference on Image Processing (ICIP)*, Kobe, Japan, October 24-28, pp. 135-139, (1999)
- [18] H. Saito, T. Kanade, Shape Reconstruction in Projective Grid Space from a Large Number of Images, *Computer Vision and Pattern Recognition Workshops (CVPR99)*, (1999)
- [19] S.J. Zhang, C. Wang, X.Q. Shao, and W. Wu, DreamWorld: CUDA-Accelerated Real-Time 3D Modeling System. Proceedings of the *IEEE International Conference on Virtual Environments, Human-Computer Interfaces, and Measurement Systems (VECIMS)*, Hong Kong, China, May 11-13, pp.168-173 (2009)
- [20] W. Wu, Z. Yang, K. Nahrstedt, G. Kurillo, and R. Bajcsy, Towards Multi-site Collaboration in Tele-immersive Environments. Proceedings of the *15th International Conference on Multimedia*, ACM, New York, NY, USA, pp. 767-770 (2007)
- [21] P. Kauff, O. Schreer, R. Tanger, Virtual Team User Environments-A Mixed Reality Approach for Immersive Tele-Collaboration. Proceedings of the *International Workshop on Immersive Telepresence (ITP)*. pp. 1-4 (2002)
- [22] D. Reidsma, R. op den Akker, R. Rienks, et al: Virtual Meeting Rooms: From Observation to Simulation. *Journal of AI & Society*, Vol.22, Issue.2, pp.133-144 (2007)
- [23] A. Nijholt, J. Zwiers, J. Peciva, The Distributed Virtual Meeting Room Exercise. *International Workshop on Multimodal Multiparty Meeting Processing*, Trento, Italy, pp. 93-99 (2005)
- [24] B.R. Duffy, G.M.P. O'Hare, A.G. Campbell, NEXUS: Fusing the Real & the Virtual. *International Conference on Computer As A Tool (EUROCON'05)*, Serbia & Montenegro, Belgrade, November, pp.22-24 (2005)
- [25] C-Lab, Germany, [Http://www.c-lab.de/jartoolkit/](http://www.c-lab.de/jartoolkit/)
- [26] M. Dragone, T. Holz, G.M.P. O'Hare, Mixing Robotic Realities. Proceedings of the *11th International Conference on Intelligent User Interfaces (IUI06)*, January 29-February 01, Sydney, Australia (2006)
- [27] B.R. Duffy, M. Dragone, G.M.P. O'Hare, et al: Fusing Realities in Human-Robot Social Interaction. Proceedings of *37th International Symposium on Robotics ISR/Robotik*, Munich, Germany, May 15-17 (2006)
- [28] M. Dragone, T. Holz, B.R. Duffy, et al: Social Situated Agents in Virtual, Real and Mixed Reality Environments. Proc. of the *Int. Conf. on Intelligent Virtual Agents (IVA '2005)*, Kos, Greece, September (2005)
- [29] M. Courgeon, J.C. Martin, C. Jacquemin, User's Gestural Exploration of Different Virtual Agents' Expressive Profiles. Proc. of *7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 08)*, Padgham, Parkes, Müller and Parsons (eds.), May 12-16, Estoril, Portugal (2008)
- [30] H.J.W. Spoelder, L. Renambot, D. Germans, et al: Man Multi-Agent Interaction in VR: a Case Study with RoboCup. In *IEEE Virtual Reality 2000* (poster), March (2000)
- [31] O. Michel, Webots: Symbiosis between Virtual and Real Mobile Robots, in: J.-C. Heuding (Ed.), Proceedings of the *First International Conference on Virtual Worlds (VW'98)*, in: LNCS/AI, vol. 1434, Springer, pp. 254-263 (1998)
- [32] F. Papadopoulos, K. Dautenhahn, W. C. Ho, and M. Walters, AIBOcom: Designing Robot Enhanced Human-human Remote Communication Technology, the *2nd International Conference on Kansei Engineering and Emotional Research 2010*, Paris, France (2010)
- [33] S.O. Adalgeirsson and C. Breazeal, MeBot: A robotic platform for socially embodied telepresence, Proceedings of the *5th ACM/IEEE International Conference on Human-Robot Interaction*, Osaka, Japan, pp.15-22 (2010)
- [34] M. L. Walters, K. Dautenhahn, R.Boekhorst, K. L.Koay, D. S. Syrdal, and C. Nehaniv, An Empirical Framework for Humanrobot Proxemics. in *AISB2009: Proceedings of the Symposium on New Frontiers in Human-Robot Interaction*. Edinburgh, Scotland, pp. 144-149 (2009)
- [35] J. Saunders, C.L. Nehaniv, K. Dautenhahn, and A. Alissandrakis, Self-Imitation and Environmental Scaffolding for Robot Teaching. *International Journal of Advanced Robotics Systems*, Vol. 4, Issue 1. pp. 109-124 (2007)

Shujun Zhang received her Ph.D degree on computer science from Ocean University of China in 2007. She completed her post-doctoral research in State Key Laboratory of Virtual Reality Technology and Systems, Beihang University from 2007 to 2009 and now is a lecturer in Qingdao University of Science & Technology. Her major field of study is virtual reality, including: 1) image-based modeling, V-R interaction and immersion; 2) computer graphics and images.

Wan Ching Ho received his first BSc and PhD degrees from University of Hertfordshire in year 2002 and 2005 respectively. The title of his PhD thesis is "Computational Memory Architectures for Autobiographic and Narrative Virtual Agents". Since 2006 he has been a full-time postdoc research fellow in the Adaptive Systems Research Group in the same university. His research mainly focuses on developing control architectures for narrative and autobiographic virtual agents, high-level cognitive robots in human-robot interaction contexts. He has been working in various EU funded projects, including ELVIS, eCircus (Framework 6) and LIREC (Framework 7).

The Research of Image Encryption Algorithm Based on Chaos Cellular Automata

Shuiping Zhang

Jiangxi University of Science and Technology Ganzhou City, Jiangxi Province
zhsp@mail.jxust.cn

Huijune Luo

Jiangxi University of Science and Technology Ganzhou City, Jiangxi Province
hjl@mail.jxust.cn

Abstract—The Research presents an image encryption algorithm which bases on chaotic cellular automata. This algorithm makes use of features that extreme sensitivity of chaotic system to initial conditions, the cellular automaton with a high degree of parallel processing. The encryption algorithm uses two-dimensional chaotic system to Encrypt image, Then establish a cellular automaton model on the initial encrypted image. Encryption key of this algorithm is made up of the initial value by the two-dimensional chaotic systems, parameters, two-dimensional cellular automata local evolution rules f and iterations n . Experimental results shows that the algorithm has features that high efficiency, better security, sensitivity to the key and so on.

Keywords— Cellular Automata; Logistic map; Image encryption; Chaos Matrix

I. INTRODUCTION

With the development and popularization of network information technology, the network information security technologies are paid more attention by society and many scholars. Information encryption and information hidden are important parts of information security technology. And image encryption technology is a critical one in it. Because images have the features of abundant data, high redundancy, low network transmission. The traditional information encryption methods have not suited for encryption of digital image, also they are not accorded the development trend of modern cryptography. With the development of the technology of image processing and modern cryptography, Many researchers spend a lot of manpower and material resources on the technology of image encryption. In recent years, due to the development of chaos theory and the further research of Cellular Automata, it realizes plenty of new technologies and new algorithms based on chaos or CA, which promote image encryption.

Because chaos system is very sensible to the initial condition and has the aperiodicity of movement track, so it adapts to data information encryption very much. Britain mathematician Matthes^[1] is the first person who applied chaos theory for the research of encryption communication technology. Chaos is a complex

dynamics behavior of some special character, it has features of extreme sensibility to initial condition, no regularity of movement track, randomness, etc. Along with chaos theory's deep research, The based on chaos theory's encryption technology also obtained the fast development. its application also no longer limits to the text secret communication domain. It is turning toward the direction development which are the multimedia information secret communication and the modern cryptology system.

Cellular Automata is a dynamics system, which defines on a CA space composed of scatter, limited state CA, and follows certain local rule, evolves on scatter time dimension^[2]. Because CA has the simplicity of the inherent component units, locality of effect among units, high parallelism of information process, and complex globality, so it makes CA fits to apply in cryptology^[3]. In 1986, Wolfram firstly put forward CA encryption technology^[2]. Since the following few years, many academicians have taken relative research on CA encryption. After several dozens years, many researchers conduct the research in cellular automaton's related characteristic. Simultaneously they also discovered that cellular automaton's related characteristic and the application of it need further studying, particularly the application domain of the two-dimensional cellular automaton and the multi-dimensional cellular automaton even are broader.

After further research on the basic of the research of chaos system to extreme sensibility to initial condition and complexity of CA evolution behavior, this paper put forward an image encrypt algorithm which combine two dimension chaos system with two dimension CA. Use two dimensional chaotic system to produce two dimensional chaotic matrix, put the chaotic matrix XOR the original image, and get a processed image. Finally, establish a model of two-dimensional cellular automaton on the processed image, then encrypt image through the evolution of cellular automata, to realize the image encryption method with higher security and better algorithm efficiency

II. TWO-DIMENSION CELLULAR AUTOMATA

The chaos system is one dynamics system about complex non-linear. It have extreme sensitivity to the initial condition, the mixing property, diffusibility and so on. The characteristics of chaos system are in keeping with the cryptology character. It has provided the new mentality and method for the development of cryptology which the application of these characteristics of chaos system have been in the secret communication domain. There are many different between the chaos system and the cryptographic system still. And the most greatly different of they is the chaos definition in continually the sequel, but cryptographic system's operation only limits in the finite field. Therefore the chaos system cannot use in the cryptographic system as a digitization method in directly; Simultaneously there are not all chaos systems that suits to design the password. The chaos system has the unique characteristic, just like ergodicity, in randomness, boundedness, fractal dimension and so on compares with the other complicated systems, these characteristics also precisely is deciding the relation about the chaos system and the cryptology.

The usual chaos dynamic systems have Logistic map, Lonenr, Rossler, Chen's System etc^[6], and Logistic map system is a scattered chaos system, which is simple and practical, its one-dimensional form researches more, but it is also an ordinary chaos encryption system, so it is hard to guarantee the security^[4]. But the two-dimensional Logistic mapping system is complex compared the one-dimensional one, if it is used in cryptographic system, the cryptographic system's security can be to obtain more safeguards. Therefore While we can use two-dimension Logistic map, select different parameters to produce two dimension chaos point sets^[5], besides to make it discretization properly and use at encryption process, by doing it the encryption algorithm is two dimension^[4]. This paper mainly applies it to the initial process(scrambling image pixel). According to one dimension Logistic map, it can define two dimension Logistic map:

$$\begin{cases} x_{n+1} = 4\mu_1 x_n(1-x_n) + g_1(x_n, y_n) \\ y_{n+1} = 4\mu_2 y_n(1-y_n) + g_2(x_n, y_n) \end{cases} \quad (1)$$

$g_1 = \gamma y_n$ and $g_2 = \gamma x_n$ are two dimension Logistic map which have first-order coupling item^[6], the mapping type is:

$$\begin{cases} x_{n+1} = 4\mu_1 x_n(1-x_n) + \gamma y_n \\ y_{n+1} = 4\mu_2 y_n(1-y_n) + \gamma x_n \end{cases} \quad (2)$$

Controls parameter μ_1 , μ_2 and γ decide the dynamics behavior. when $\mu \geq 0.89$, $\gamma = 0.1$, initial point $(x_0, y_0) = (0.10, 0.11)$, the system is chaos^[6]. Selecting the control parameter $\mu_1 = \mu_2 = 0.90$, $\gamma = 0.1$, initial parameter $(x_0, y_0) = (0.10, 0.11)$, this time we can get the two-dimensional Logistic mapping phase diagram as shown in Figure 1. The set of points of the system produced is the chaos from the chart.

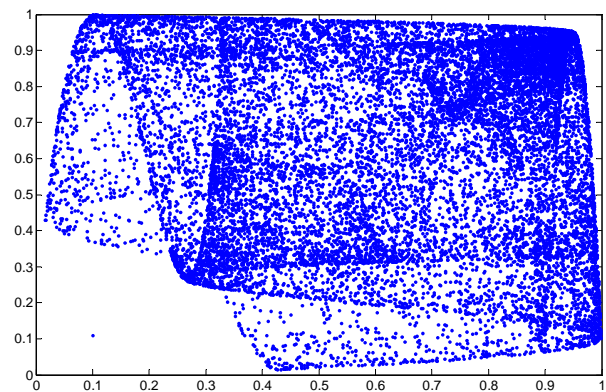


Figure 1. Two-dimensional Logistic mapping phase diagram

III. IMAGE ENCRYPTION ALGORITHM OF CHAOS CELLULAR AUTOMATA

As one kind of complicated system's model, cellular automaton's composition is actually quite simple, it is composed of four mainly basic parts, they are respectively the structure cell, the structure cell space, the neighbors of cell and the transformation rule. Therefore it is thought that the cellular automaton may also is constituted by a structure cell space and the transformation rule defined in this cell space. Actually, the transformation rule is a condition transfer function, it is a dynamics function that it determines this structure cell condition next time according to the current condition of the structure cell and the neighbors.

The Cellular Automata's evolution are mainly relied on its local transform rule, each CA's current state s_i^{t+1} is determined by the $2r$ neighbor cellular automata with it and center CA's preceding time, $s_i^{t+1} = f(s_{i-r}^t, \dots, s_i^t, \dots, s_{i+r}^t)$, f is cellular automata's local transform rule, r is cellular automata's radius, t is time. All cells in the structure cell space rely on the local transformation rule to change at the same time. sometimes, the combination of all spatial structure cell condition in the cell space is a structure cell configuration. According to cellular automata space grid distribution, CA can divide into one-dimension, two-dimension and Multiple-dimension cellular automata. The elementary cellular automata with one-dimension radius $r = 1$, CA's states is 0 or 1 more widely by researched. The two-dimensional cellular automaton is that the transformation rule distributes on grid points in the two-dimensional Euclidean plane. Its application is most widespread, the model is the two-dimensional cellular automaton model most. Comparing with one-dimension CA, two-dimension CA is much more complicated no matter from evolution behavior or neighbor definition, Fig2, Fig3 and Fig4 are Von. Neumann model, Moore model and Expansion model of Moore neighbor of the two-dimension CA^[2].

Supposing the state sets of two-dimension is $s = \{0, 1\}$ two states, neighbor is Von. Neumann model, it compose

of a center CA and four surrounding CAs, which corresponding local state evolution rule is:

$$s_{i,j}^{t+1} = f(s_{i-1,j}^t, s_{i,j-1}^t, s_{i,j}^t, s_{i,j+1}^t, s_{i+1,j}^t) \quad (3)$$

Following the rule of up to down ,left to right ,two-dimension CA can translate into one- dimension CA, from Fig1,it corresponding to one dimension CA which radius $r = 2$,according to the computer method of elementary CA, different evolution rules corresponding to different rule numbers , it is :

$$R = \sum_{i=0}^{2^k-1} s_i \times 2^i \quad (4)$$

k is cellular number , s_i is the state of the i cellular^[2,7]

,when $k = 5$, its whole rules are 2^{2^5} . By analyzing the relations between one-dimension and two-dimension CA's rule number ,and their neighbor radius, it can infer that with the increase of neighbor radius, rule number's space shows exponential rise, so making the rule space parameterize, it benefits the research of CA behavior characteristic. As one evolving dynamics system, cellular automaton whether be a evolved language, or a evolved behavior, it displayed the extremely complex characteristic, particularly, cellular automaton's evolved behavior displays the very complex multiplicity. It mainly discusses the two dimension CA's evolution behavior characteristic. To any transform function, it defines a corresponding parameter value λ ,

$$\lambda = (m^{2^{r+1}} - n_q) / m^{2^{r+1}} \quad (5)$$

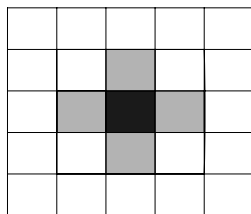


Figure 2. Von. Neumann model

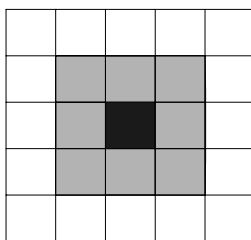


Figure 3. Moore model

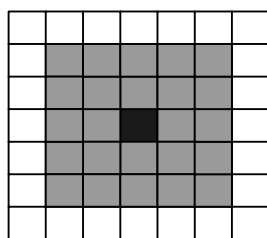


Figure 4. Expansion mole of Moore

m is the state number of state set $S = \{s_1, s_2, \dots, s_m\}$; r is neighbor radius ; n_q is the number of all outputs which are 0, parameter λ is close to the CA' evolution behavior. Through Wolfram's research shows that, with the parameter λ changing from 0 to 1. CA evolution behavior presents different states. About elementary cellular automata, it can divide into different types on it. When λ is between 0 and 0.1, the CA evolution behavior performance is the uniform state, named the point attractor state. This kind of CA is stationary; When λ is between the 0.1 and 0.3, the evolution behavior of CA is periodic state, and it is cycle track, this is the cycle model of CA; When λ is between 0.3 to 0.6 ,the evolution behavior of CA is a fairly complex structure, then the evolution behavior is a complex local structure . Such is the complex CA; When λ is greater than 0.6, the evolution behavior of CA has no complex structure, and the performance of the chaotic behavior of the model, a completely random, chaotic state, that is chaotic attractor, corresponding to the chaotic CA model Machine; Most of the evolution of CA under the same rules, the initial conditions on the evolution behavior of the state has little effect. As λ increasing, the evolution behavior of CA gradually becomes random and complex ,and behavior evolved from the initial state into ordered structure. When it satisfies the cellular automata rule with $\gamma = 0.5$, the uncertainty (Information Entropy)of evolve into producing one serial is the maximum, this part of rule numbers of corresponding CA apply on encryption ,while it can be classified to complexity in elementary CA. The structure cell unit's independence decided this high parallelism of the cellular automaton information processing, that has provided more advantageous condition for some algorithm which its performance required high encryption algorithm. specially the two-dimensional cellular automaton's unique feature are similar to image data characteristic, this has provided more widespread theory method of graphic processing and the encryption to the cellular automaton. looking from the bulk properties, the one-dimensional cellular automaton and the two-dimensional cellular automaton are very similar, but many evolution systems involve the structure cell shape and the extremely complex boundary condition, therefore the two-dimensional cellular automaton's evolved behavior must be much more complex than the one-dimensional. The reversible cellular automaton is that some original state the cellular automaton arrived at another condition by the partial regulation f_0 and evolution n times, the another condition can returns to the original state by partial regulation f_1 and evolution n times, then we said that the cellular automaton of rule f_0 and the cellular automaton of the rule f_1 is mutually reversible cellular automaton, what is called the invertibility cellular automaton. According to the invertibility of CA global function, CA can be divided into invertible CA and irreversible CA^[8], among them, the invertibility of one dimension CA can judge by CA evolution rule, while two dimensions CA cannot^[9]. References[9] gives a method of constructing an irreversible two dimension based on four one dimension

invertible CA rule, on this basis, further study found that it can also use one dimension CA rule with a radius $r = 1$ and number 0 elementary CA with the XOR method to construct a two-dimension Von. Neumann model reversible CA transform rule . it's construct formula :

$$F = f_1 \oplus f_0. \tag{6}$$

f_1 is one dimension reversible CA of radius $r = 1$, f_0 is number 0 one dimension CA , F is two dimension Von. Neumann model reversible, and the generating process of evolution rule's local map of number F CA shows as Fig5. With one dimension reversible CA of rule number 85 of radius $r = 1$ and number 0 CA , using the formula (6) to calculate the two dimension Von. Neumann model's corresponding rule number is $R=85899345$ reversible CA, construct a pair of two dimension reversible CA with rule number 16711935 , among them, the rule number $R=85899345$ two dimension Von. Neumann model CA's local rule(following the sequence of up, left, right, down) shows as Table I .According to (5) formula, it can work out the parameter $\lambda=0.5$ of number 85899345 two dimension CA . For further researching this two-dimensional cellular automaton's evolution behavior, we found that its evolution behavior is same as the primary 85 cellular automaton's evolution behavior, they are one kind of cycle cellular automaton, and it is also one kind of reversible two-dimensional cellular automaton. In theory, the CA space is infinitely extended ,but on practical applications, it needs to confirm the CA's boundary condition, there are three usual CA boundary conditions : constant boundary,periodic boundary, reflective boundary, sometimes stochastic pattern is also possible to use, it real-time has the stochastic values in the boundary.. It adopts cycle type boundary in this paper . the parameter $\lambda=0.5$ of number 85899345 two dimension CA . For further researching this two-dimensional cellular automaton's evolution behavior, we found that its evolution behavior is same as the primary 85 cellular automaton's evolution behavior, they are one kind of cycle cellular automaton, and it is also one kind of reversible two-dimensional cellular automaton. In theory, the CA space is infinitely extended ,but on practical applications, it needs to confirm the CA's boundary condition, there are three usual CA boundary conditions : constant boundary,periodic boundary, reflective boundary, sometimes stochastic pattern is also possible to use, it real-time has the stochastic values in the boundary.. It adopts cycle type boundary in this paper .

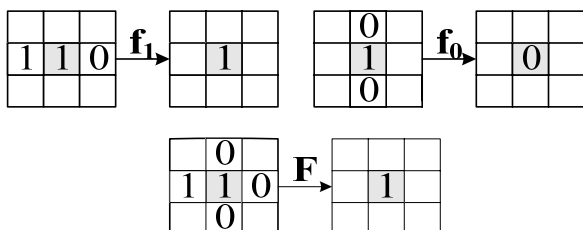


Figure 5. Two dimension Von. Neumann CA local map

TABLE I. NUMBER 85899345'S TWO DIMENSION CA RULE TABLE

t	t+1	t	t+1
0000	1	1000	1
0001	1	1001	1
0010	0	1010	0
0011	0	1011	0
0100	1	1100	1
0101	1	1101	1
0110	0	1110	0
0111	0	1111	0
1000	1	1100	1
1001	1	1101	1
1010	0	1110	0
1011	0	1111	0
1100	1	1100	1
1101	1	1101	1
1110	0	1110	0
1111	0	1111	0

IV. IMAGE ENCRYPTION ALGORITHM OF CHAOS CELLULAR AUTOMATA

A. Encryption Principle

The image encrypt algorithm's principle based on chaos CA: firstly, to use the two dimension Logistic map of first-order coupling form mentioned before, to make sure the initial value x_0, y_0 , to select proper value of μ_1, μ_2 and γ , according to the size of encrypting image matrix , to produce two chaos series of $\{x_0, x_1, \dots, x_n\}$ and $\{y_0, y_1, \dots, y_n\}$,to form chaos matrix which has the same size as image matrix, to make XOR between chaos matrix and original image matrix , to get Initialization processing , then build the model of two reversible CA , to ensure evolution rule f , to iterate n times following the rule , the end , to obtain the image after it is encrypted .

B. process of encryption and decryption algorithm

Following the above encryption principle , to set the size $M \times N$ of image A to be encrypted , the process of encryption algorithm as follows .

(1)According to initial value and system parameters(the value of initial Encryption Keys x_0, y_0, μ_1, μ_2 and γ) , after two dimension Logistic map of first-order coupling form ,it generates two chaos series $\{x_k, k=1,2,\dots,(M \times N)/2\}$ and $\{y_k, k=1,2,\dots,(M \times N)/2\}$ at length of $M \times N/2$ on the basis of image size .

(2)Use the two chaos series to generate a chaos matrix B with the same size as image matrix (it's odd-number rows are replaced by the pre-N of chaos series $\{x_k, k=1,2,\dots,(M \times N)/2\}$ as a row of chaos matrix , and its even-number rows are replaced by the pre-N

chaos series $\{x_k, k=1,2,\dots,(M \times N)/2\}$, so it gets a chaos matrix B with same size as image .).

(3)To take XOR on bit between each element of chaos B and each pixel value, to generate the image C by initial process.

(4)To build a two dimension reversible CA model on image C, and to ensure the neighbor model and local transform rule f , the to iterate n times following evolution rule f , in the end to get image Z after it is encrypted, the encryption is :

$$k = (x_0, y_0, \mu_1, \mu_2, \gamma, f_1, n)$$

The process of decryption algorithm is the inverse of encryption algorithm, the difference is the decryption Key of decryption:

$$k' = (x_0, y_0, \mu_1, \mu_2, \gamma, f_2, n)$$

V. EXPERIMENTAL RESULT AND ANALYSIS

This paper takes gray scale image which is named Lena, and size of 128x128 for an example, Fig6 is original picture. It takes $(x_0, y_0) = (0.10, 0.11)$, $\mu_1 = \mu_2 = 0.89$, $\gamma = 0.1$ as the initial values and system parameters of two dimension chaos system(2), uses the Von. Neumann neighbor model, and CA boundary condition takes the cycle boundary, and takes two dimension reversible CA of rule number $r=858993459$ with iteration number $n=25$, then to encrypt based on the above encryption algorithm. Fig7 is the encrypted image.

A. analyze encryption key space

In encryption key attack method, Brute-force attack and encryption key analysis is the most basic and commonly used method. From the perspective of cryptology, the quality of encryption algorithm depends on the size of encryption key space, so the size of encryption key space determines the algorithm safety. In this algorithm, the encryption key is divided into two parts, one part is two initial conditions (x_0, y_0) and three control parameters $(\mu_1 = \mu_2 = 0.89, \gamma = 0.1)$ of chaos system ;



Figure 6. original image

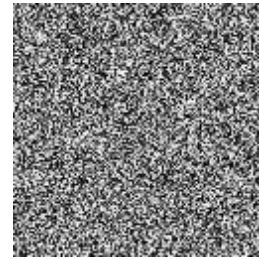


Figure 7. encrypted image

another part is local transformation rule f and the iteration number n of two dimension CA, this paper selects two initial conditions and parameters of an accuracy is 10^{-2} , the rule space of Von. Neumann model two dimension CA is 2^{32} , and the iteration number is $n = 25$, so the encryption key space is :

$$n \times 10^2 \times 10^2 \times 10^2 \times 10^2 \times 10^2 \times 2^{32} = n \times 4.295 \times 10^{19}$$

With the increase of iteration number n and the expansion of accuracy parameters, the encryption key space will continue to increase, It is impossible to use brute-force attack, the algorithm is effective safe.

B. Sensitivity analysis of encryption key

In order to test the sensitivity of the encryption key, this paper selects the $\mu_1 = 0.8900000001$ and the other parameters unchanged to decrypt the encrypted image, to get as shown in Fig9. Fig8 is the correct decryption result. So it can see that the same secret key figure cannot obtain the correct decryption image in the case of small changes. Therefore, this encryption algorithm has the ability of good resistance to differential analysis.

C. Statistical feature analysis

Fig10 is the original image's histogram, Fig11 is the histogram which uses the encryption algorithm after 25 iterations to the original image, it can be seen from the diagram, the histogram has significant changes after encrypted, the image pixels tend to uniform distribution after 25 iterations encrypted, so it is well covered up the distribution of the image before encryption, and has strong random. So the encryption algorithm has good diffusion to the image pixels.

D. Interdependency Analysis of neighbor pixels

To further analyses on interdependency of the original image and the encrypted image's neighbor pixels, this paper randomly selects pixels ^[10-11] from 1000 pairs of adjacent images (including horizontal, vertical and diagonal direction), by the formula:

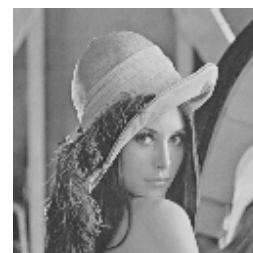


Figure 8. correct decryption result

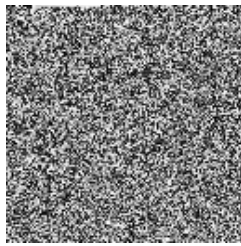


Figure 9. wrong decryption result

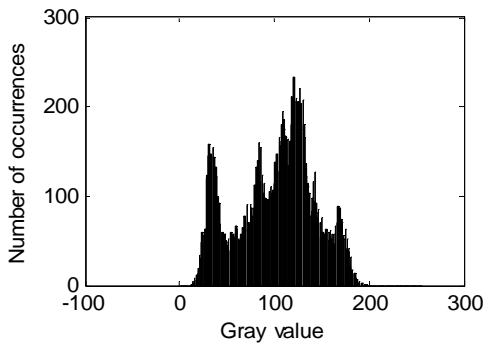


Figure 10. Original histogram

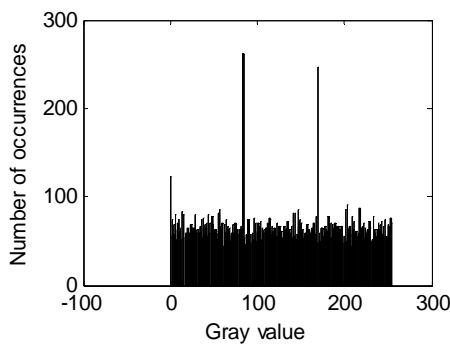


Figure 11. encrypted original histogram

$$r_{xy} = \frac{\text{cov}(x, y)}{\sqrt{D(x)} \cdot \sqrt{D(y)}} \quad (7)$$

It calculates the interdependency between two graphs, where x, y are two adjacent pixels of the gray value, r_{xy} is the correlation coefficient, $\text{cov}(x, y)$ is the covariance, $D(x), D(y)$ is variance [11]. Table II lists the neighbor pixel's interdependency of the original image and the encrypted image (iteration 25) in the horizontal, vertical, diagonal direction, Fig12 is the original image's neighbor pixels in the horizontal direction interdependency, and Fig13 is encrypted image neighbor pixels after 25 times iteration in the horizontal direction interdependency. Fig14 and Fig15 are respectively original images and encrypted images Statistical correlation charts in the vertical direction. Fig16 and Fig17 are respectively original images and encrypted images statistical correlation charts in the diagonal direction. It can see from Table II ,the correlation coefficient closes to 1 in the original image neighbor pixels, but it is almost 0 in

the encrypted image. The encrypted image neighbor pixel interdependency is greatly reduced, neighbor pixels have been largely irrelevant. Seen from Fig12, Fig13, Fig14, Fig15, Fig16 and Fig17, So it proves that the statistical characteristics of the original image have spread to the encrypted image.

E. Analysis of average change in gray value and image similarity

To further analyze the safety extend of encrypted image, the paper introduces the average change in gray value and image similarity extend analysis [12]. The formula is:

$$GAVE(G, C) = \frac{\sum_{i=1}^M \sum_{j=1}^N |G_{ij} - C_{ij}|}{M \times N} \quad (8)$$

$$XSD(G, C, \alpha, \beta) = 1 - \frac{\sum_{i=1}^M \sum_{j=1}^N [C_{ij} - G_{ij}]^2}{\sum_{i=1}^M \sum_{j=1}^N G_{ij}^2} \quad (9)$$

In formula G_{ij} is the gray value in the original image's the i row j column pixel with size $M \times N$, C_{ij} is the gray value in the encrypted image's the i row j column pixel, in the formula (9), the α, β are two integers, and $0 \leq \alpha < M - 1, 0 \leq \beta < N - 1$, in the image encryption algorithm, the more regular gray value changes with encrypted image and original image, the better to encryption and security, and the best case that the average

TABLE II. THE INTERDEPENDENCY BETWEEN ORIGINAL IMAGE AND THE ENCRYPTED IMAGE NEIGHBOR PIXELS

Direction	original image	encrypted image
neighbor pixels in horizontal direction	0.8058	-0.0280
neighbor pixels in vertical direction	0.8979	0.0622
neighbor pixels in diagonal direction	0.7732	-0.0475

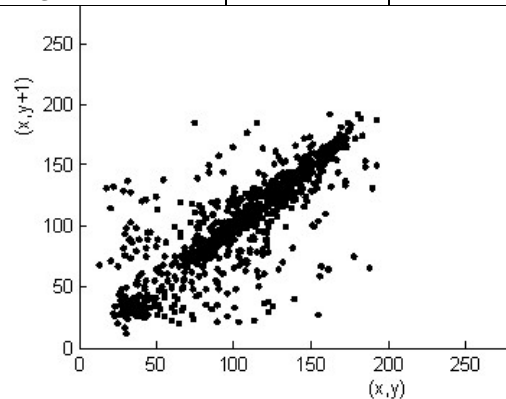


Figure 12. original image's interdependency in horizontal direction

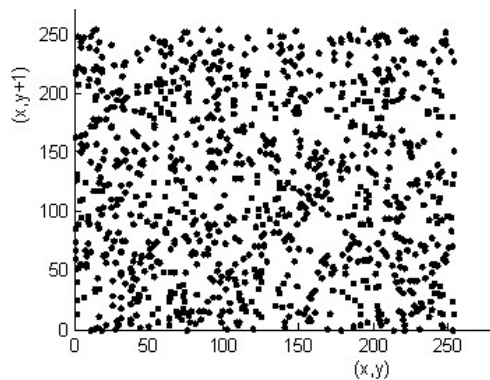


Figure 13. encrypted image's interdependency in horizontal direction

gray change value is as half as image gray^[12]. The similarity between encrypted image and the original image is also an important index to judge the safety of image encryption algorithm. If two images are completely alike, the similarity is 1, and $\alpha = \beta = 0$. The greater change between the encrypted image and original image, the smaller with the similarity of two images, the smaller similarity between the encrypted image and original image, the higher with the safety^[12]. With the formula (8) we can calculate the encrypted image and original image's gray average change value is 73.0762 in the experimental algorithm; According to the formula (9) to calculate two images' similarity, in this paper the encrypted image and original image's similarity is 0.3513. It can be seen from the results, although the gray average change value of the encrypted image and the original image could not be the best, but the similarity of the encrypted image and the original image is small, so the algorithm is safe and effective.

F. Analysis of Algorithm complexity

The time and space complexity of the algorithm are one of the factors what determines the performance of the algorithm. The algorithm's calculated time focuses on generating chaos matrix, XOR operation and cellular automata iterations. Setting an image with size of $M \times N$, each operating time is f , the number of iterations is n , so the time of generating the chaos matrix is $T_1 = (M \times N) \times t$; XOR operation time is T_2 , as each pixel translates into an 8-bit binary to operate XOR, the time is $T_2 = 8 \times (M \times N) \times t$; The cellular automaton iteration time is T_3 , because the value of each pixel is translated into 8 bits, so the time is $T_3 = (n + 2) \times (M \times N) \times t$; The algorithm to complete the encryption process needs $T = ((n + 3) \times 8 + 1) \times (M \times N) \times t$, so the time complexity of this algorithm is $O(n \times M \times N)$. If the computer takes full use of cellular automata parallelism, to take the parallel algorithm, its time capability will be better; if n is given, the time complexity is $O(M \times N)$. On lattice complexity, the computation process of this algorithm requires two matrixes with lattice size $M \times N$ and a matrix with size $M \times N \times 8$, the algorithm in the lattice complexity is $O(M \times N)$; It can be seen that the algorithm is good in lattice complexity and

time complexities, so the algorithm is a better capability algorithm.

VI. CONCLUSION

The Research puts forwards an image encryption algorithm based on chaos cellular automata, to take full use of the two-dimensional chaotic system's extreme sensitivity to initial conditions and the features with parallelism, complexity and randomness of two-dimensional reversible cellular automata, to make the encrypted key, encrypted image and original image much more have complexity, randomness and unpredictability. Besides to take the two-dimensional chaotic system parameters and initial value, the rule f of two-dimensional reversible cellular automaton and the iteration number n as the encrypted key, greatly to expand the encrypted key space. It is effective to resist exhaustion attack and differential analysis. Through experimental test and analysis of the algorithm, it shows that the algorithm has high safety and efficiency.

REFERENCES

- [1] Matthes R. On the derivation of a Chaotic Encryption algorithm [J]. *Cryptologia*, 1989.XIII (1) :29-42
- [2] Wolfram s. Cryptography with cellular automata [A]. *Advance in cryptology: Crypto's 85 ProceeDings*, Lecture Note in Computer science [C]. Heidelberg: Springer ,1986,429-432.
- [3] Zhang Chuanwu, Shen Ye Qiao, Qi-Cong Peng. Cellular Automata Encryption Based on the reverse iteration [J]. *Journal of Computers*, 2004,27 (1) :125-129.
- [4] Gao Shan, Xu Songyuan, Sun Baiyu so, the encryption process based on chaos theory research [J]. *Automation Technology and Applications*, 2001, (6) :13-16.
- [5] WANG Xing-yuan. *Complex nonlinear systems Chaotic* [M]. Beijing: Electronic Industry Press, 2003.
- [6] Chen Yongqiang, Sun Huaning, based on the number of two-dimensional chaotic map image encryption algorithm[J], *Wuhan Polytechnic University*, 2004,12, Vol.23, No.4 :45-47. In number: TP391.
- [7] Zhong-Jun Wang, Neng-Chao Wang, Feng Fei, Tian Wufeng, the evolution of cellular automata behavior[J], the computer application, 2007,8, Vol.24, No.8 ,38-41, in number: TP391.41
- [8] Ping Ping, Zhou Yao, Zhang, Feng-Yu Liu, reversible cellular automata encryption technology[J], *Communications*, 2008,5, Vol.29, No.5 ,26-32, in number: TP309.7.
- [9] Zhu Baoping, Liang Zhou, Yu-Feng Liu, cellular automata based on public key cryptography research[J], *Nanjing University*, 2007,10, Vol.31, No.5 :612-616, in number: TP309.7.
- [10] Chen G, Mao YB, Chui C K. A symmetric image encryption scheme based on 3D chaotic cat Maps [J]. *International Journal of Bifurcation and Chaos*, 2004,14 (10) :3163-3624.
- [11] Visualization of chaotic, Huang Huiqing, two-dimensional chaotic system based on digital image encryption algorithm[J], *Shantou University (Natural Science)*, 2009,2, Vol.24, No.1 :56-61, in number: TP391.
- [12] Wang winding Ran, Chun-Xia Wang, Zhan Xinsheng, an image encryption algorithm performance assessment method[J], *Computer* ,2006,10-3 :321-314, in number: TP309 + .2.

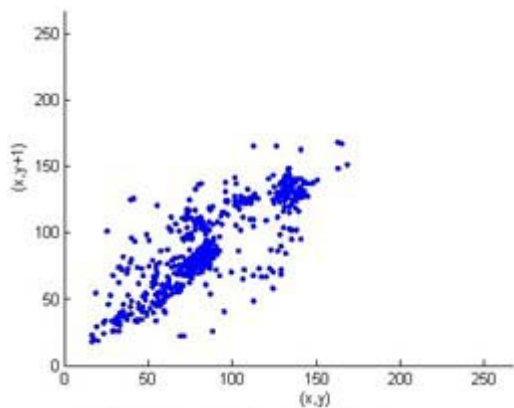


Figure 14. original image's interdependency in vertical direction

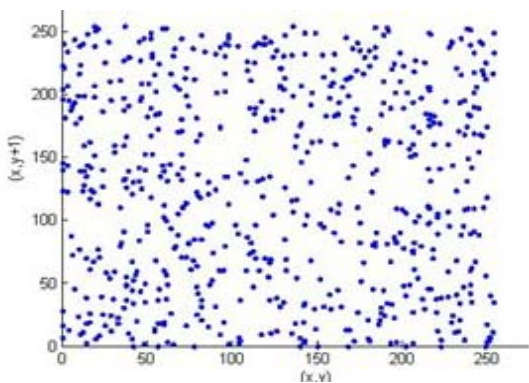


Figure 15. encrypted image's interdependency in vertical direction

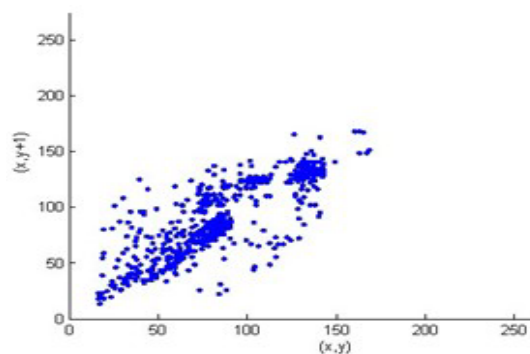


Figure 16. original image's interdependency in diagonal direction

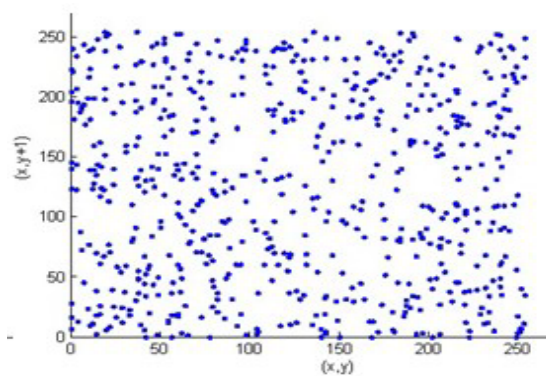


Figure 17. encrypted image's interdependency in diagonal direction

Improved MFCC Feature Extraction Combining Symmetric ICA Algorithm for Robust Speech Recognition

Huan Zhao, Kai Zhao, He Liu

School of Information Science and Engineering, Hunan University, Changsha, China
Email: hzhao@hnu.edu.cn, zhaokai@hnu.edu.cn, liuhe@hnu.edu.cn

Fei Yu

Jiangsu Provincial Key Laboratory of Computer Information Processing Technology, Suzhou, China
Email: hunanyufei@126.com

Abstract—Independent component analysis (ICA), instead of the traditional discrete cosine transform (DCT), is often used to project log Mel spectrum in robust speech feature extraction. The paper proposed using symmetric orthogonalization in ICA for projecting log Mel spectrum into a new feature space as a substitute in extracting speech features to solve the problem of cumulative error and unequal weights that deflation orthogonalization brings, so as to improve the robustness of speech recognition systems, and increase the efficiency of estimation at the same time. Furthermore, the paper studied the nonlinearities of the objective function in ICA and their coefficients, tested them in all kinds of environments, finding that they influenced the recognition rate greatly in speech recognition systems, and applied a new coefficient in the proposed method. Experiments based on HMM and Aurora-2 speech corpus suggested that the new method was superior to deflation-based ICA and MFCC.

Index Terms—independent component analysis, speech feature extraction, speech recognition

I. INTRODUCTION

Speech feature extraction has been a key focus in robust speech recognition research^[1]. Selecting appropriate features guarantees the good performance of a speech recognition system. Among a large amount of methods for speech feature extraction, the ones based on spectrum are widely used, especially Mel frequency cepstral coefficients (MFCC). Although many new methods for feature extraction are proposed constantly, such as non-stationary feature extraction^[2], Gabor analysis and tensor factorization based feature extraction^[3], etc, MFCC is still the most important

method for speech feature extraction in state-of-the-art automatic speech recognition systems.

Because the feature space by DCT is not dependent on real speech data directly, MFCC performs poor in noisy environment. Data-driven feature space transformations are highly adaptable to real speech data, and will achieve better results than DCT in a practical environment. Principle component analysis (PCA), linear discriminant analysis (LDA) and independent component analysis (ICA) are frequently-used data-driven linear transformations. These transformations replace DCT in MFCC procedure to transform the feature space of logarithmic spectrum for new speech features. On the basis of the principle of minimum reconstruction error, PCA projects spectral coefficients onto the direction of maximum variance. ICA performs feature transformation based on the hypothesis of statistical independence of independent components, expecting to find the original structure of speech features.

Independent component analysis has become an important method in statistics, and makes significant progress especially in the field of blind source separation^[4]. Recently ICA draws more and more attention in speech feature extraction. FastICA method is widely used because of its high efficiency^[5], mainly in speech feature extraction when used in speech recognition. When estimating many independent components, there are two decorrelation modes in FastICA: deflation (serial) and symmetric (parallel) orthogonalization method^[6]. The paper discussed two different methods in speech feature extraction, and talked about the nonlinearities of objective function in FastICA and their coefficients.

Feature transformation is a common method in speech feature extraction, projecting the feature space in order to achieve decorrelation^[7], dimensionality reduction and noise reduction. There are two main categories^[8]: linear feature transformations, such as DCT, PCA, LDA, and ICA, etc.; nonlinear feature transformations, such as nonlinear principal component analysis (NPCA), nonlinear discriminant analysis (NLDA), nonlinear

National Science Foundation of China (Grant NO. 61173106), the Key Program of Hunan Provincial Natural Science Foundation of China (Grant No.10JJ2046), the Planned Science and Technology Key Project of Hunan Province, China (Grant No.2010GK2002).

Corresponding author: Huan Zhao(email: hzhao@hnu.edu.cn)

independent component analysis (NICA) and so on. Reference [8] applied PCA, LDA, ICA and nonlinear LDA in a phone recognition task using TIMIT, and compared the results of the different speech features. Reference [9] extracted the correlation information of subspace of phones using PCA in order to extracting speech features. Reference [10] transformed some different speech features using LDA, and reduced the recognition error rate efficiently. Taking the computational complexity and accuracy into account, linear feature transformations methods are commonly used, and applied after getting the Log Mel spectrum. DCT is a non-data related transformation, so it can't adapt to the characteristics of the actual data, and achieves only partial decorrelation^[11]. LDA determines complexly and is sensitive to the mismatch of SNR of training and testing set. On the basis of the principle of minimum reconstruction error, PCA projects spectral coefficients onto the direction of maximum variance. ICA regards the inputted multidimensional data as a linear combination of independent components and reestimates the original independent components according to some objective, in order to obtain the physical structure and formation of these components^[12]. After pre-emphasis, frame windowing, FFT, Mel filtering and logarithms, feature coefficients are gotten. MFCC, PCA features, and ICA features are gotten when applying DCT, PCA and ICA respectively to feature coefficients. Based on deflation and symmetric decorrelation categories, ICA features can be classified into deflation ICA features (ICA_DEFL) and symmetrical ICA features (ICA_SYMM). In the experiments the paper compared the influence of four different features on robustness and accuracy of automatic speech recognition systems. The following of the paper first introduced the ICA principle and described the feature extraction method using ICA. At the same time the paper researched the influence of nonlinearities of objective function and their coefficients on automatic speech recognition systems, and then tested them to verify the performance. Finally, the paper discussed and summarized the experimental results.

II. FEATURE EXTRACTION BASED ON SYMMETRIC ICA

A. The Principle of ICA

Independent component analysis (ICA) is a method which finds internal factors or components from multivariate statistical data^[12], looking for both statistically independent and non-Gaussian components. ICA is used in blind source separation at the earliest, but recently also applied to feature extraction gradually. In reference [13] the author used ICA to replace the Fourier transform. In reference [11] ICA was applied to log Mel spectrum. Assuming observed random variables x_1, x_2, \dots, x_n , each of which is a linear combination of another n random variables s_1, s_2, \dots, s_n :

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \dots + a_{in}s_n, i = 1, \dots, n \quad (1)$$

In (1), $a_{ij}, i, j = 1, \dots, n$ are real coefficients, assuming all s_i are statistically independent. Only random variables x_i can be observed, a_{ij} and s_i must be estimated just by x_i . Eq. (1) can be show using matrix as $x = A * s$. Random vector x represents mixed vector, s represents independent components, and A represents a matrix which is composed of a_{ij} . To obtain independent components, a demixed matrix W should be computed:

$$u = W * x \quad (2)$$

Where W is the inverse matrix of matrix A , and u is an estimate of s .

B. Feature Extraction Based on Symmetric ICA

According to different principles, there are various methods to estimate W in ICA, such as maximizing the nongaussianity method, maximum likelihood estimation method and minimizing the mutual information method, etc. An important method of maximizing the nongaussianity methods is FastICA. When estimating multiple independent components using FastICA, they can be estimated one by one using deflation orthogonalization algorithm one by one. Each time one vector w_i is initialized, updated, orthogonalized, and normalized until it converges. Independent components also can be estimated using symmetric orthogonalization method. Every w_i is iterated firstly, and then all w_i are orthogonalized using a special way.

Deflation (serial) orthogonalization method and symmetric (parallel) orthogonalization method computes W respectively as Fig. 1:

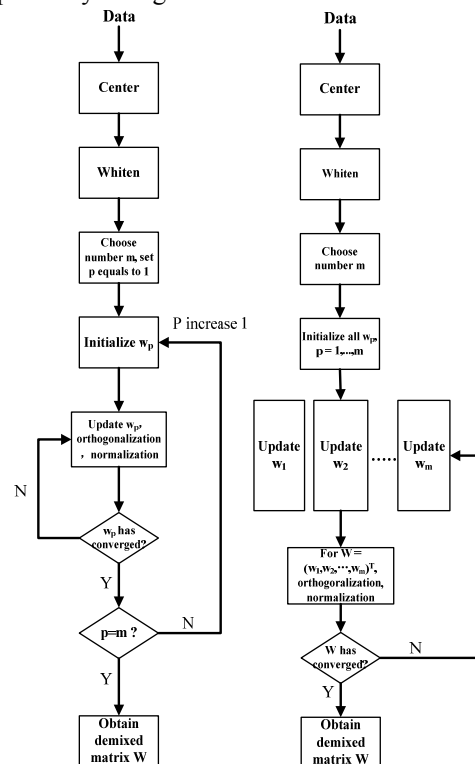


Figure 1. The deflation orthogonalization of ICA and symmetric orthogonalization of ICA.

The difference between the two methods lies in calculating the demixed matrix W in different ways. The

former calculates each component of W one by one, updates and orthogonalizes them using (3) and (7) respectively until they converge while the other calculates them in parallel, updates and orthogonalizes them using (3) and (8) respectively until they converge.

$$w_p = E\{zg(w_p^T z)\} - E\{g'(w_p^T z)\}w_p \quad (3)$$

Where g can be (4), (5) or (6)

$$g_1(y) = \tanh(a_1 * y) \quad (4)$$

$$g_2(y) = y * \exp(-a_2 * u^2 / 2) \quad (5)$$

$$g_3(y) = y^3 \quad (6)$$

$$w_p \leftarrow w_p - \sum_{j=1}^{p-1} (w_p^T w_j) w_j \quad (7)$$

$$W = (WW^T)^{-1/2} W,$$

$$\text{where } W = (w_1, w_2, \dots, w_p)^T \quad (8)$$

There are some deficiencies in deflation orthogonalization method: the error of firstly estimated components will be accumulated so as to influence the estimate of following components which is brought by the orthogonalization. Independent components can be calculated in parallel using symmetric orthogonalization method to solve the above problem, meanwhile the time of calculating W can be shortened sharply. In following experiments, we extracted speech features using the two methods, compared the influence on automatic speech recognition systems which they brought, and discussed the advantages and disadvantages of them.

C. Nonlinearities and their coefficients

The statistical properties of ICA (such as consistence, asymptotic variance, robustness) depend on the selection of objective functions. In objective functions, the non-quadratic functions G are very important, and provide high-level information in the form of expectation $E\{b_i^T x\}$. In actual algorithms, this is equivalent to choosing the derivative of G , nonlinearities g . In the following of the paper, G and g were both referred as nonlinearities. Reference [6] proved that the optimal non-quadratic functions are in the following form:

$$G_{opt}(y) = |y|^a, a < 2 \quad (9)$$

However, the problem of the above functions is that they are not derivable in origin when $a \leq 1$ and this leads to numerical optimization problem. Reference [6] indicated that: (1) A good general-purpose function is $G(y) = \log(\cosh(a_1 * y))$, where $1 \leq a_1 \leq 2$ and a_1 is a real number; (2) When independent components are super-Gaussian or robustness is very important,

$G(y) = -\exp(-y^2 / 2)$ may work well; (3) Only when independent components are sub-Gaussian and there are no outliers, kurtosis is proper. Three functions are as follows:

$$G_1(y) = \frac{1}{a_1} \log(\cosh(a_1 * y)) \quad (10)$$

$$G_2(y) = -\exp(-a_2 * y^2 / 2) \quad (11)$$

$$G_3(y) = \frac{1}{4} y^4 \quad (12)$$

In reference [4] the author did experiments in brain imaging and image feature extraction using the above nonlinearities, however, no one had researched the role of nonlinearities in speech feature extraction. The paper would discuss the selection of nonlinearities and their coefficients. When a is equal to 1, $G_{opt}(y) = |y|$ is not derivable in origin and always replaced by $G(y) = \log(\cosh(a_1 * y))$, where a_1 is a real number. The two functions are as Fig. 2:

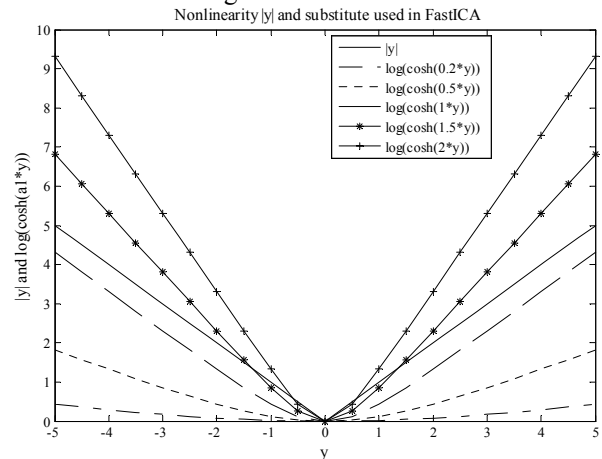


Figure 2. $G_{opt}(y) = |y|$ and $G(y) = \log(\cosh(a_1 * y))$

As can be seen from Fig. 2, the closer to 1 the coefficient a_1 is, the closer to optimal function $|y|$ the function $\log(\cosh(a_1 * y))$ is, while the curve is steeper. Reference [12] proposed that $G(y)$ should not grow too fast with $|y|$, otherwise it would rely on some observations that are far from the origin. We must make a compromise between the approximation of accuracy and the function's smoothness. The selection of a_1 will rely to specific applications and is not absolute. When using $\log(\cosh(a_1 * y))$ as the function to extract speech features, we should test a_1 and find the optimal value.

D. Features Selection

As for DCT, the reserved first several coefficients can be treated as speech features^[13]. For ICA, feature selection can be processed according to the L2-norm of ICA basis. ICA basis refers to the column vector of the inverse of ICA demixed matrix. The L2-norm of basis represents the contribution of the basis to the whole signal. The bigger the value of the norm is, the more the

contribution is. The L2-norm of ICA bases are shown in Fig. 3.

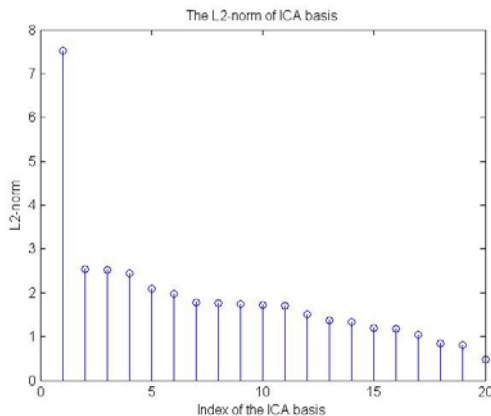


Figure 3. The L2 norm of bases of ICA

Therefore the most important N basis can be chosen according to the bigger value of L2-norm value of bases and used in speech feature recognition.

III. EXPERIMENTS

A. Results of Four Kinds of Features

The paper used speech recognition development toolkit HTK of Cambridge University to build a speech recognition system based on HMM which was used to assess the new features based on symmetric orthogonalization ICA and other features. Moreover, the paper compared the new features with MFCC, PCA features, and the features based on deflation orthogonalization ICA. The experiments were performed on the Aurora-2 corpus, sampled to 8 kHz. The 80 speakers (consisting of 40 male and 40 female) from the training subset were selected. The training set consisted of 100 male and 100 female speeches which were clean. There were 40 male and 40 female speeches in each noise environment and SNR in testing set. There were 8 kinds of practical environments: airport, babble, car exhibition, restaurant, street, subway and train. The SNR were divided into 7 classes: -5db, 0db, 5db, 10db, 15db, 20db and clean. The total number of training and testing set was 4680.

The speech signal was divided into frames of 32 ms in length with an overlap of 10ms between frames. Pre-emphasis and Hamming window were applied to each frame first. Then FFT was performed then to extract the spectrum. Mel filter bank analysis with 20 channels was processed for each frame. Logarithm operation was performed following FFT. These coefficients were transformed by the DCT to get the traditional MFCC features. By using the PCA-based, deflation ICA-based and symmetric ICA-based transformation, PCA features and ICA features were obtained. The last two ICA-based features were denoted as ICA_DEFL and ICA_SYMM respectively. The final feature vector consisted of 13 components with first-order deltas and second-order deltas. There were 39 components in the final features in each case. The experiments extracted the four features in

Matlab and tested them in a speech recognition system built using HTK. The results were as Fig. 4:

As can be seen from Fig. 4, when SNR was high or in clean condition, the performance of ICA_SYMM was almost the same as the other three features, while features based on ICA were superior to others in low SNR. Table 1 and Fig. 5 showed the average recognition rate of the four features in 8 noise environments. From them two, we can find the two features based on ICA were both superior to MFCC greatly. The recognition rate of ICA_SYMM feature was about 6.17% higher to MFCC. At the same time, ICA-based features were better than PCA-based feature, excepting that ICA_DEFL was lower than PCA in exhibition environment. The ICA_SYMM feature was better than ICA_DEFL feature, proving the accuracy of discussion of the two orthogonalization of ICA.

B. The Nonlinearities and Their Coefficients

In FastICA, the common nonlinearities of objective functions were (10), (11) and (12). Reference [14] proposed using rational nonlinearities as substitutes of the three common nonlinearities to reduce the mass of computation. However, experiments proved that it didn't work well. In our experiments, ICA_SYMM features were computed using the three nonlinearities respectively. The results were as Table 2 and Fig. 6.

As can be seen from Fig. 6, $G_1(y)$ was superior to $G_2(y)$ and $G_3(y)$ when computing average speech recognition rate in all 8 noise environments.

The coefficient a_1 would affect the property of $G_1(y) = \frac{1}{a_1} \log(\cosh(a_1 * y))$. The experiments researched the influence of a_1 on recognition rate when its value was between 0 and 2. Results of the experiments were as Table 3: in the street, car, airport, exhibition, restaurant and train environment, the maximum average recognition rate was reached when a_1 equaled 0.2; in babble environment, the maximum average recognition rate 67.32% was reached when a_1 equaled 1.1, but it was only 0.65% higher than that gotten when a_1 equaled 0.2; in subway environment, the maximum average recognition rate 66.03% was reached when a_1 equaled 0.6, and it was 1.87% higher than that gotten when a_1 equaled 0.6. As can be seen, when a_1 equaled 0.2, excellent performance could be gotten in most noise environments, and the value performed quite well in other special environments too. The experiments proved that the value 0.2 of a_1 is a reliable empirical value in most situations. The value 0.2 of a_1 could be used to improve the performance of automatic speech recognition systems when extracting speech features using ICA-based method to extract speech features in noise environments.

In Fig. 7 below: the horizontal axis was the value of a_1 , from 0.1 to 2, and the interval was 0.1; the vertical axis was the average recognition rate in noise environments. In (a) and (b), the value of a_1 influenced the average speech recognition rate in a certain trend, and the best performance was reached when the value was 0.2.

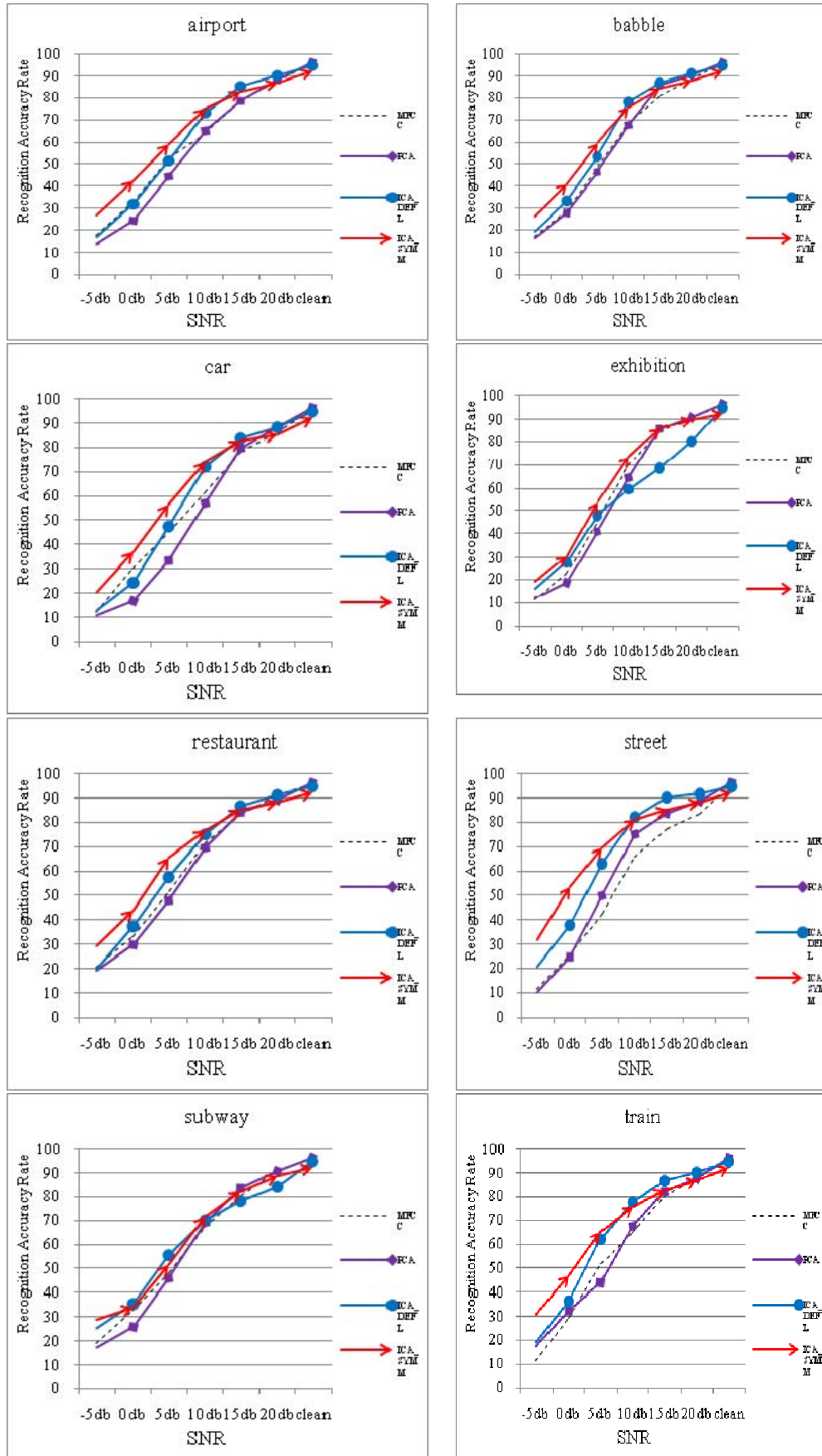


Figure 4. Recognition accuracy rate in each environment and each SNR (%)

TABLE I. AVERAGE RECOGNITION RATE OF THE FOUR FEATURES IN 8 ENVIRONMENTS (%)

	street	babble	car	train	subway	restaurant	airport	exhibition
MFCC	57.61	61.18	58.41	60.17	62.14	64.01	61.34	60.06
PCA	61.29	61.51	54.63	61.03	61.29	62.40	58.73	58.46
ICA_DEFL	68.59	65.18	60.54	66.72	63.42	66.14	63.37	56.49
ICA_SYMM	71.78	66.67	64.11	68.69	64.16	68.80	66.51	63.57

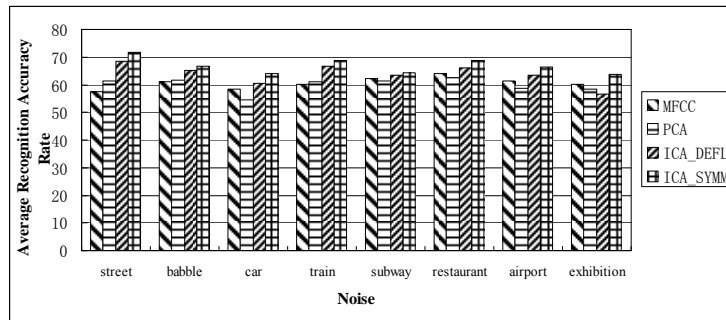


Figure 5. Average recognition rate of the four features in 8 environments (%)

TABLE II. AVERAGE RECOGNITION RATE OF THE THREE NONLINEARITIES IN 8 ENVIRONMENTS (%)

	airport	babble	car	exhibition	restaurant	street	subway	train
$G_2(y)$	56.25	57.63	51.56	54.38	58.49	60.62	54.28	55.29
$G_3(y)$	58.02	60.41	54.60	56.36	58.87	66.44	58.12	60.36
$G_1(y)$	66.51	66.67	64.11	63.57	68.80	71.78	64.16	68.69

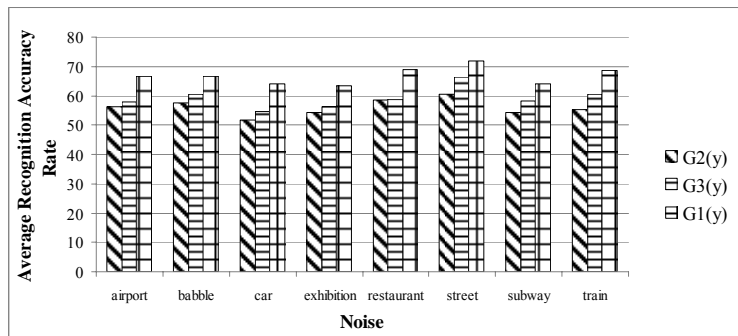
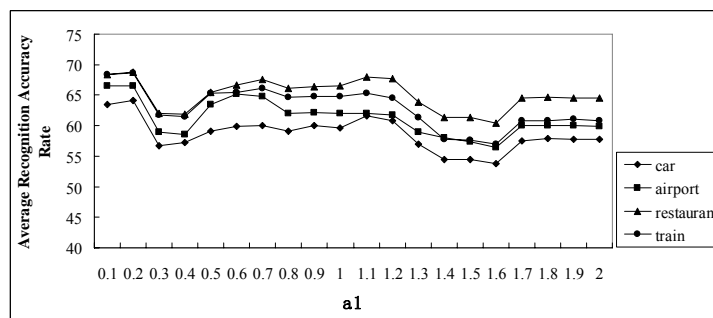
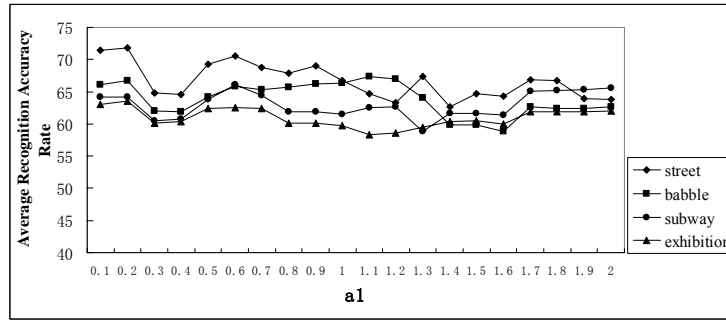


Figure 6. Average recognition rate of the three nonlinearities in 8 environments (%)



(a)



(b)

Figure 7. Average recognition rate influenced by different value of a1 in 8 environments (%)

TABLE III. THE AVERAGE RECOGNITION RATE IN DIFFERENT VALUES OF A1 IN 8 ENVIRONMENTS (%)

	street	babble	car	subway	airport	exhibition	restaurant	train
0.1	71.46	66.13	63.47	64.21	66.51	63.04	68.37	68.32
0.2	71.78	66.67	64.11	64.16	66.51	63.57	68.80	68.69
0.3	64.85	61.97	56.75	60.54	58.97	60.06	61.98	61.71
0.4	64.58	61.87	57.18	60.75	58.56	60.32	61.87	61.49
0.5	69.23	64.22	59.16	63.80	63.53	62.36	65.39	65.29
0.6	70.51	65.82	59.85	66.03	65.18	62.57	66.62	65.50
0.7	68.74	65.28	60.01	64.38	64.75	62.35	67.58	66.14
0.8	67.88	65.76	59.10	61.92	62.03	60.11	66.14	64.70
0.9	69.01	66.19	60.01	61.93	62.19	60.12	66.41	64.86
1.0	66.72	66.30	59.69	61.55	61.98	59.74	66.56	64.86
1.1	64.75	67.32	61.66	62.56	62.03	58.35	68.00	65.28
1.2	63.35	66.99	60.86	62.67	61.76	58.56	67.68	64.53
1.3	67.42	64.06	56.97	58.78	59.00	59.47	63.90	61.34
1.4	62.62	59.89	54.46	61.66	57.98	60.37	61.28	57.71
1.5	64.75	59.84	54.51	61.66	57.39	60.53	61.34	57.60
1.6	64.37	58.88	53.82	61.33	56.48	59.95	60.37	56.96
1.7	66.88	62.67	57.55	65.02	60.06	61.87	64.48	60.75
1.8	66.72	62.41	57.87	65.18	60.01	61.87	64.64	60.86
1.9	63.95	62.46	57.82	65.34	60.06	61.88	64.54	61.02
2.0	63.84	62.61	57.82	65.60	59.95	61.98	64.54	60.86

IV. CONCLUSION

The paper proposed using symmetric orthogonalization ICA-based method to extract speech features, and verified the new features in 8 different kinds of noise environments. The experiments proved that the average recognition rate of the new features was 6.17% higher than that of MFCC features, especially excellent in low SNRs. The new method got better performance than deflation orthogonalization ICA-based method and MFCC, and improved the robustness of the speech recognition system and the efficiency of estimation of ICA. The nonlinearities of objective function in ICA and their coefficients had a great impact on recognition

accuracy rate, when $G_1(y) = \frac{1}{a1} \log(\cosh(a1 * y))$ and a1 =

0.2 it got the best performance in general. Because the demixed matrix of ICA in the new method was calculated offline, it could be calculated firstly before used in extracting speech features in fact which would save much time for estimation. From the above, we can see that the new method improved the average recognition rate but didn't strength the complexity of computation, so it is possible for the new method to replace MFCC as a popular method extracting speech features in the future.

REFERENCES

- [1] U. Shrawankar and V. Thakare. "Feature Extraction for a Speech Recognition System in Noisy Environment: A Study," Computer Engineering and Applications (ICCEA), 2010 Second International Conference on. 2010.
- [2] Z. Tuske, P. Golik, R. Schluter, and F.R. Drepper. "Non-stationary feature extraction for automatic speech recognition," Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. 2011.
- [3] W. Qiang, Z. Liqing, and S. Guangchuan, "Robust Multifactor Speech Feature Extraction Based on Gabor Analysis," Audio, Speech, and Language Processing, IEEE Transactions on, 2011. vol 19(4), pp. 927-936, 2011.
- [4] H. Hsieh, J. Chien, K. Shinoda, and S. Furui. "Independent component analysis for noisy speech recognition," Acoustics, Speech and Signal Processing, (ICASSP). IEEE International Conference on. 2009.
- [5] E. Ollila, "The Deflation-Based FastICA Estimator: Statistical Analysis Revisited," Signal Processing, IEEE Transactions on, 2010. vol 58(3), pp. 1527-1541, 2011.
- [6] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," Neural Networks, IEEE Transactions on, 1999. vol 10(3), pp. 626-634, 1999.
- [7] X. Zou, P. Jancovic, and M. Kokuer, "On the Effectiveness of the ICA-based signal representation in non-Gaussian Noise," Icsip: 2008 9th International Conference on Signal Processing, vols 1-5, pp. 1-4, 2008.
- [8] P. Somervuo, "Experiments with linear and nonlinear feature transformations in HMM based phone recognition," 2003 Ieee International Conference on Acoustics, Speech, and Signal Processing, vol I, pp. 52-55, 2003.
- [9] P. Hyunsin, T. Takiguchi, and Y. Arika. "Integration of Phoneme-Subspaces Using ICA for Speech Feature Extraction and Recognition," Hands-Free Speech Communication and Microphone Arrays, 2008. HSCMA 2008. 2008.
- [10] R. Schluter, A. Zolnay, and H. Ney. "Feature combination using linear discriminant analysis and its pitfalls," INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, INTERSPEECH 2006 - ICSLP, September 17, 2006 - September 21, 2006. 2006. Pittsburgh, PA, United states: DUMMY PUBID.
- [11] L. Potamitis, N. Fakotakis, and G. Kokkinakis, "Independent component analysis applied to feature extraction for robust automatic speech recognition," Electronics Letters, vol 36(23) pp. 1977-1978, 2000.
- [12] A. Hyvärinen, J. Karhunen, and E. Oja, Independent component analysis. vol. 26, 2001.
- [13] J.H. Lee, H.Y. Jung, T.W. Lee, and S.Y. Lee, "Speech feature extraction using independent component analysis," 2000 Ieee International Conference on Acoustics, Speech, and Signal Processing, vols I-Vi, pp.1631-1634, 2000.
- [14] P. Tichavský, Z. Koldovský, and E. Oja, "Speed and accuracy enhancement of linear ICA techniques using rational nonlinear functions," Independent Component Analysis and Signal Separation, pp. 285-292, 2007.



Huan Zhao is a professor at the School of Information Science and Engineering, Hunan University. She obtained her B.Sc. degree, M.S. degree and Ph.D. in Computer Science and Technology at Hunan University in 1989, 2004 and 2010, respectively. Her current research interests include speech information processing, embedded system design and embedded speech recognition. She served as visiting scholar at the University of California, San Diego (UCSD), USA during the period of March 2008 to September 2008. The visiting scholarship was appointed and sponsored by the China Scholarship Council (CSC). Prof. Zhao is a Senior Member of China Computer Federation, Governing of Hunan Computer Society, China and China Education Ministry Steering Committee Member of Computer Education on Arts. She has published more than 40 papers and 9 books.



Kai Zhao received his B.Sc. degree in Computer Science and Technology at the school of Computer and Communication, Hunan University, P. R. China in 2009. Currently, he is a M.S. candidate of Hunan University, P. R. China. His current research interests include speech feature extraction and speech recognition.



He Liu received her B.Sc. degree in Electronic Information Engineering at the School of Computer and Electronic Engineering, Hunan University of Commerce, P. R. China in 2009. Currently, she is a M.S. candidate of Hunan University, P. R. China. Her current research interests include digital signal processing, speech information processing and feature extraction.

A Watermarking Technique based on the Frequency Domain

Huang-Chi Chen

Department of Electrical Engineering, I-Shou University, Kaohsiung City, Taiwan
 Department of Management Information System, Far East University, Tainan County, Taiwan
 E-mail: hchen46@cc.feu.edu.tw

Yu-Wen Chang

Department of Electronic Communication Engineering, National Kaohsiung Marine University,
 Kaohsiung City, Taiwan
 E-mail: may@webmail.nkmu.edu.tw

Rey-Chue Hwang

Department of Electrical Engineering, I-Shou University, Kaohsiung City, Taiwan
 E-mail: rchwang@isu.edu.tw

Abstract—A watermarking technique based on the frequency domain is presented in this paper. The one of the basic demands for the robustness in the watermarking mechanism should be able to dispute the JPEG attack since the JPEG is a usually file format for transmitting the digital content on the network. Thus, the proposed algorithm can be used to resist the JPEG attack and avoid the some weaknesses of JPEG quantification. And, the information of the original host image and watermark are not needed in the extracting process. In addition, two important but conflicting parameters are adopted to trade-off the qualities between the watermarked image and the retrieve watermark. The experimental results have demonstrated that the proposed scheme has satisfied the basic requirements of watermarking such as robustness and imperceptible.

Index Terms—watermark, JPEG, frequency domain.

I. INTRODUCTION

Digital watermarking is an effective and common technique to protect intellectual property rights [1-5]. Most of watermarking techniques [6-8], the watermark will be embedded into the frequency domain instead of the spatial domain for the robustness of the watermarking mechanism. The DCT coefficients of the host image will be modified in order to hide the watermark by using the embedding rule and the information of the watermark. For the DCT coefficients which are modified in the high-frequency region, the affect for the quality of the image is slight. However, the watermark will be damaged for the signal processing. On the other hand, although the information is hidden into the low-frequency region can avoid JPEG compression attacks. But the quality of the host image would be destroyed seriously. Therefore, the watermark had been chosen to embed into the middle-frequency region mostly [3-11]. But sometime, the original host image and watermark are needed in the extracting process [4].

Until now, the various watermarking techniques are proposed and have been highly effective on academic aspect of watermarking. A VQ-based digital watermarking scheme with the codebook expansion method is presented in [11-12]. Some of these techniques are simple and effective, but the embedded watermark is fragile. Recently, the methods using the soft computing are introduced to optimum the scaling parameters intelligently in order to trade-off the qualities between the host image and the retrieved watermark [6, 14]. However, these mechanisms are not designed as oblivious. A hybrid watermark technique based on Genetic Algorithm (GA) and Particle Swarm optimization (PSO) is developed in [15]. Although two complementary watermarks are embedded simultaneously to provide higher detection response, each watermark could resist a specific class of attacks. Until the correspondence between the communication and the data hiding is proposed, the digital watermarking technique is regarded as a communication problem. Therefore, the error correcting codes are used to raise the robustness of the watermark [11, 16-17]. For the oblivious watermarking technique [8-10, 19], the watermark would be extracted without the original image by comparing and modifying the DCT coefficients of each two blocks. Thus, the data of the media didn't need to save twice.

The clearly fact is that almost all of the digital contents transmitting over the network belong to the compressed file for the consideration of the frequency bandwidth. The JPEG is a popular compacting format of file to travel the media over the internet. Nevertheless, two similar DCT coefficients will be quantified to the same rank in JPEG quantification. Thus, this question discussed above will cause the mistake of the extracting watermark from the watermarked image with lossy JPEG compression. Furthermore, the robustness and imperceptibility are the

two important factors for the watermarking scheme, but they are opposite to achieve [18].

In this paper, a modified algorithm is presented to improve the defect of the JPEG quantification in order to reduce the bit error rate (BER) of the retrieved watermark. Addition, two parameters are regarded as the controlling factors. They are used to adjust the value of the DCT coefficient in order to trade-off the qualities between the watermarked image and retrieve watermark. Moreover, the proposed algorithm is design as a blind mechanism. Thus, the original image and watermark are not needed for extracting watermark.

II. THE BASIC OF CONCEPT

A. JPEG Quantization

It is well known that JPEG is a most commonly lossy compression standard for the digital contents traveling over the internet. To achieve the robustness for resisting the JPEG compression attacks, one watermarking scheme will be designed with the quantization algorithm of JPEG. Now discuss as follow. The overall block process of JPEG encoder and decoder is shown in Figure 1.

First, the original image was partitioned into blocks of size 8×8 without overlap. Second, these blocks must be DCT transformed from the spatial domain to the frequency domain. Third, the quantization is performed as follow:

$$b''_{u,v} = \text{round}(b_{u,v} / Q_{u,v}), \text{ for } 0 \leq u, v < 8 \quad (1)$$

where $b_{u,v}$ (resp. $Q_{u,v}$) denotes the DCT coefficient of the transform image block (resp. the quantity coefficient of the quantization table) at the position (u, v) . $b''_{u,v}$ represents the quantization rank. Next, after the encoding with the variable-length codes, the information in JPEG compression format will be transmitted over the channel. The de-compression process reverses the procedure of how the data are compressed. The quantity $b'_{u,v}$ is retrieved by the de-quantization as follow:

$$b'_{u,v} = b''_{u,v} \times Q_{u,v}, \text{ for } 0 \leq u, v < 8 \quad (2)$$

However, in JPEG quantization, there is a question as shown in Table 1. The two similar DCT coefficients will be quantified to the same quantization rank. It will cause the mistake of the retrieved watermark extracting from the watermarked image with lossy compression. Thus, a modify method is proposed in this presented scheme to improve the above weakness.

B. Watermarking in the Frequency Domain

The basic idea to embed watermarking into the frequency domain is to modify the value of the DCT coefficients. It is clearly to find that the data loss of JPEG

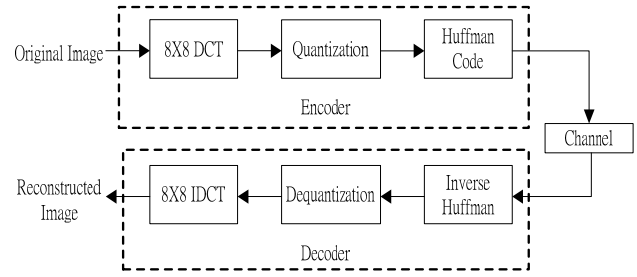


Figure 1. The whole process of JPEG codec.

TABLE 1.
Quantization of JPEG(Quantity coefficient $Q = 11$)

DCT Coefficient	Quantization Rank	DCT Coefficient (De-Quantization)
$b_1 = 8$	$b''_1 = 1$	$b'_1 = 11$
$b_2 = 14$	$b''_2 = 1$	$b'_2 = 11$

compression comes from the quantization in (1) and the de-quantization in (2) process. It can be easily found that the value of $Q_{u,v}$ at the high frequency region is higher so that the data will divided higher quantization value to product more loss. Therefore, if the watermark is embedded into high frequency region, it will be not to against the JPEG attack. On the other hand, while the watermark is hidden at low frequency region, the host image will be destroyed seriously. Therefore, the watermark will be hidden in the middle frequency region.

C. Permutation

The pixel permutation is performed on the watermark. The permutation process is defined by chaos mapping function. Let T_γ denote the permutation matrix is as follow:

$$T_\gamma = \begin{bmatrix} 1 & 1 \\ \gamma & \gamma + 1 \end{bmatrix} \quad (3)$$

For any positive integer γ , the minimum number $\rho(\gamma)$ is existing such that $T_\gamma^{\rho(\gamma)} = I_2$, where I_2 is belong the 2×2 identity matrix and the quantity $\rho(\gamma)$ is defined as the period of the permutation. It is clearly that if $a + b = \rho(\gamma)$, then $(T_\gamma^a)^b = T_\gamma^{\rho(\gamma)} = I_2$.

The permutation is performed on the pixel positions of watermark w with size $M \times M$. The permuted watermark w_p is given by

$$w_p = w_p(x', y') \quad (4)$$

$$[x' \ y']^T = T_\gamma^a \cdot [x \ y]^T \text{ mod } M$$

where $0 \leq x, y < M, a < \rho(\gamma)$. The permuted watermark can be easily retrieved by additional operation T_γ^b on w_p , where $a + b = \rho(\gamma)$.

D. Idea of Optimum

The idea of optimum is used to trade-off the qualities between the retrieve watermark and the watermarked image. Here, Δ , Γ are regarded as two scaling factors. They are used to adjust the value of the DCT coefficient in the embedding process. The intention of Δ is to reduce the variance quantity in DCT coefficient under the embedding rule. The Modification of DCT coefficient in case 2 (resp. case 3) of the embedding process is shown in Figure 2 (resp. Figure 3). The operation of Γ is as a boundary factor. In other word, the watermark bit will be give up if the variance quantity of DCT coefficient is over than the Γ . Where $\Gamma = \varepsilon \cdot Q_{u,v}$ and $\varepsilon \in [1,2]$ belongs a positive number.

III. WATERMARKING EMBEDDING

The proposed watermarking mechanism is to hide the permuted watermark into the middle frequency region in order to provide robustness. The procedure of the embedding the watermark into the host image are developed and is shown in Figure 4.

First, the permuting operation is performed on the watermark w of size 128×128 with the T_γ to obtain the permuted watermark w_p which is given by

$$w_p = w_p(x', y')$$

where

$$[x' \ y']^T = T_\gamma^a \cdot [x \ y]^T \text{ mod } 128$$

and

$$0 \leq x, y < 128, a < \rho(k).$$

On the host image f of size 256×256 , a block partition operation is adopted to product blocks of size 8×8 without overlap. The block transform of the block

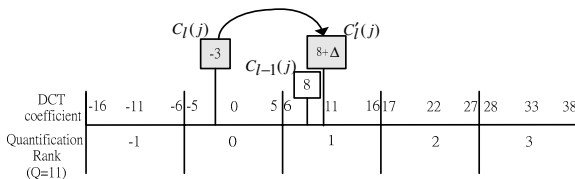


Figure 2. Modification of DCT coefficient in case 2 of the embedding process

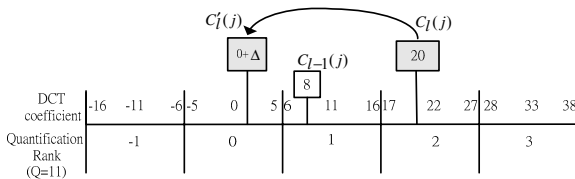


Figure 3. Modification of DCT coefficient in case 3 of the embedding process

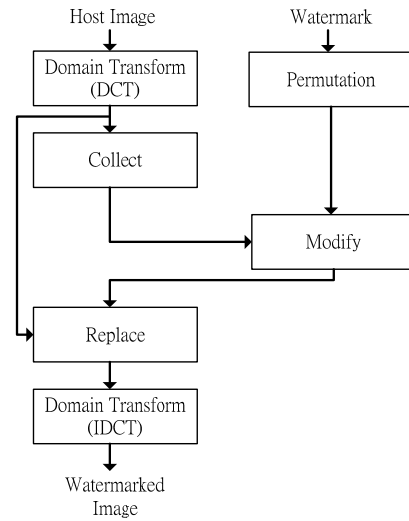


Figure 4. The process of the watermarking embedding.

image f_p using DCT is denoted as $F = \{F(u,v) | 0 \leq u,v \leq 255\}$. Where, these blocks are sequentially labeled as C_l , for $0 \leq l \leq 1023$.

To continue, the 16 elements are collecting out from the middle frequency coefficients. The modification operation for these 16 coefficients according the watermark and rules 1 as follows:

Rule 1:

- (a) If $C_{l-1}(j) \leq C_l(j)$ and $w'_j = 1$,
then $C'_l(j) = C_l(j) + \Delta$
- (b) If $C_{l-1}(j) > C_l(j)$ and $w'_j = 1$,
If $(C_{l-1}(j) + \Delta - C_l(j) < \Gamma)$,
then $C'_l(j) = C_{l-1}(j) + \Delta$
else $C'_l(j) = C_l(j)$
- (c) If $C_{l-1}(j) \leq C_l(j)$ and $w'_j = 0$,
If $\left(C_l(j) - \left[\left(\frac{C_{l-1}(j)}{Q(j)} \right) - 1 \right] \times Q(j) < \Gamma \right)$,
then $C'_l(j) = \left[\left(\frac{C_{l-1}(j)}{Q(j)} \right) - 1 \right] \times Q(j) + \Delta$
else $C'_l(j) = C_l(j)$
- (d) If $C_{l-1}(j) > C_l(j)$ and $w'_j = 0$,
If $\left(\frac{C_{l-1}(j)}{Q(j)} \right) = \left(\frac{C_l(j)}{Q(j)} \right)$,
then $C'_l(j) = \left[\left(\frac{C_{l-1}(j)}{Q(j)} \right) - 1 \right] \times Q(j) + \Delta$
else $C'_l(j) = C_l(j)$
for $l = 1, 2, \dots, 1023, j = 0, 1, \dots, 15$

where $C_l(j)$ (resp. $C_{l-1}(j)$) denotes the middle coefficient on block C_l (resp. C_{l-1}); w'_j denotes the permuted watermark bit will be embedded into the C_l ;

$Q(j)$ denotes the quantification of corresponding to the quantization table in JPEG standard. Two parameters, Δ and Γ , are used to control the value of the DCT coefficient. These 16 coefficients which had be modified are restored back into F to obtain the new frequency domain image G . After the operation of Inverse DCT, a watermarked image g is obtained.

III. WATERMARKING EXTRACTING

The proposed watermarking scheme is designed as oblivious. Therefore, the original host image and watermark are all not necessitated in the extracting process. The procedure of the extracting the watermark from the watermarked image g is presented and is given in Figure 5.

On the watermarked image g of size 256×256 , a block partition operation is adopted to product blocks of size 8×8 without overlap. The block transform using DCT is used to obtain G . Next, the 16 elements which are the same positions as in the embedding method is collected out from the middle frequency coefficients.

Retrieve the permuted watermark according the rule 2 as follow:

Rule 2:

- (a) If $B_{l-1}(j) \leq B_l(j)$ then $w'_p = 1$
- (b) If $B_{l-1}(j) > B_l(j)$ then $w'_p = 0$

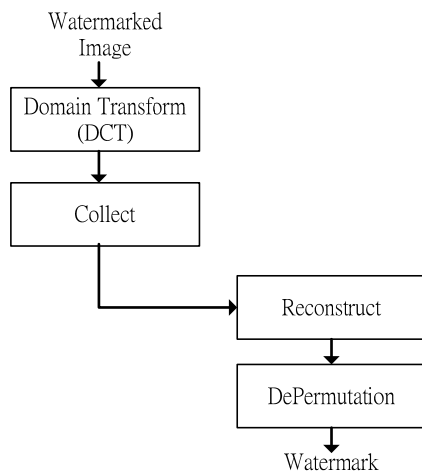


Figure 5. The process of the watermark extracting.

Finally, the retrieved watermark w' is obtained by the reversing permutation with chaos map function.

IV. EXPERIMENTAL RESULTS

To demonstrate the robustness of the proposed scheme, the algorithm has been simulated using C++ program. The host images of size 256×256 are 8-bit gray level images and the watermarks of size 128×128 are binary images. And, one watermark and five host images (i.e., Lena, F16, Pepper, Baboo, and Girl) are used to test.

The peak signal noise rate (PSNR) is used to estimate the quality between the original image and the watermarked image, which denote as f and g , respectively. The PSNR is defined as

$$PSNR = 10 \cdot \log_{10}(255^2 / MSE) \tag{5}$$

where $MSE = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} (f(x, y) - g(x, y))^2 / N^2$.

The similarity between the original watermark w and the retrieved watermark w' is measured using NC, which denote as

$$NC = \frac{\sum_{x=0}^{M-1} \sum_{y=0}^{M-1} (w(x, y) \cdot w'(x, y))}{\sum_{x=0}^{M-1} \sum_{y=0}^{M-1} w(x, y)^2} \tag{6}$$

Table 2, Table 3, and Table 4 show the quality of the watermarked images and retrieved watermarks, in which different values of the two scaling factors Γ and Δ for the different host images. Where, $\Gamma = \varepsilon \cdot Q_{u,v}$. The results prove that, in $\Gamma = 1.5 \times Q_{u,v}$, the values of PSNR are higher but the NC are lower. The values of NC are higher but the PSNR are lower in $\Gamma = 2.0 \times Q_{u,v}$. Therefore, when the $\Gamma = 1.7 \times Q_{u,v}$ and $\Delta = 0.005$, there have the balance in trade-off the qualities between the watermarked image and the retrieved watermark. The Figure 6, Figure 9, and Figure 12 are the results of the PSNR in which different values of the two scaling factors. And, Figure 7, Figure 10, and Figure 13 are the results of the BER in which different values of the two scaling factors. The values of NC under various degrees of the two scaling parameters are depicted in Figure 8, Figure 11, and Figure 14.

TABLE 2.
Quality of watermarked images and retrieved watermark in which different values of the two scaling factors Γ and Δ .

Γ	$1.5 \times Q_{u,v}$			$1.7 \times Q_{u,v}$			$2.0 \times Q_{u,v}$		
Δ	0.005	0.01	0.05	0.005	0.01	0.05	0.005	0.01	0.05
PSNR (dB)	34.15	34.11	33.54	33.8	33.74	33.16	32.90	32.83	32.16
BER (%)	6.78	6.79	6.92	5.71	5.78	5.93	4.18	4.16	4.11
NC	0.962	0.961	0.956	0.969	0.968	0.963	0.975	0.975	0.973

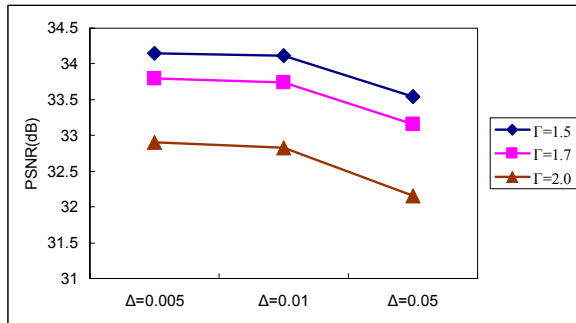


Figure 6. Quality (PSNR) of watermarked images in which different values of the two scaling factors Γ and Δ . (TABLE 2)

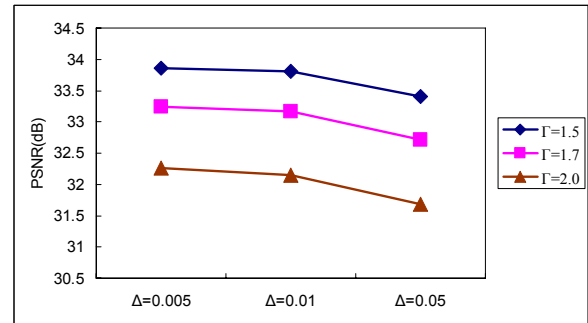


Figure 9. Quality (PSNR) of watermarked images in which different values of the two scaling factors Γ and Δ . (TABLE 3)

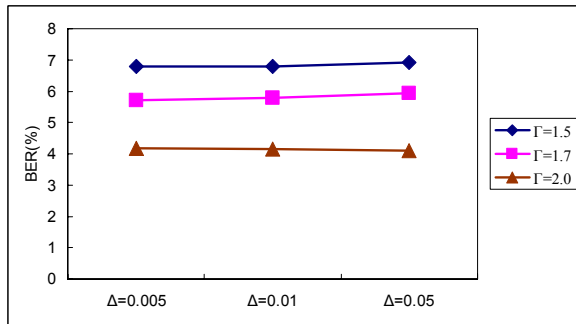


Figure 7. Quality (BER) of retrieved watermark in which different values of the two scaling factors Γ and Δ . (TABLE 2)

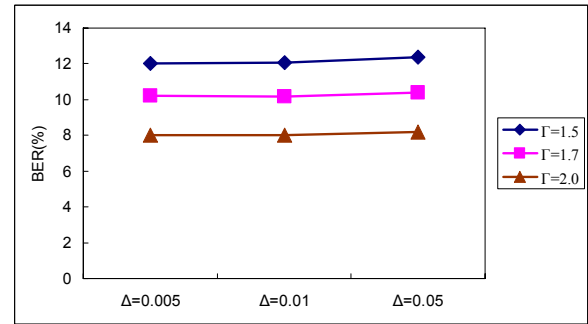


Figure 10. Quality (BER) of retrieved watermark in which different values of the two scaling factors Γ and Δ . (TABLE 3)

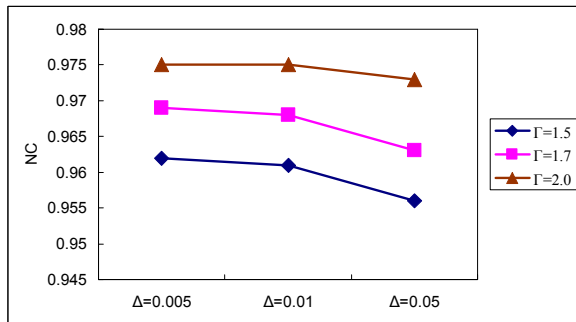


Figure 8. Quality (NC) of retrieved watermark in which different values of the two scaling factors Γ and Δ . (TABLE 2)

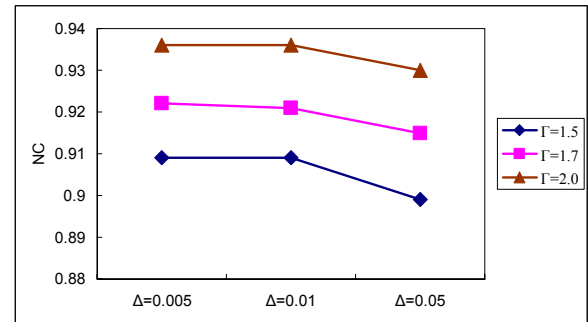


Figure 11. Quality (NC) of retrieved watermark in which different values of the two scaling factors Γ and Δ . (TABLE 3)

TABLE 3.
Quality of watermarked images and retrieved watermark in which different values of the two scaling factors Γ and Δ .



									
Γ	$1.5 \times Q_{u,v}$			$1.7 \times Q_{u,v}$			$2.0 \times Q_{u,v}$		
Δ	0.005	0.01	0.05	0.005	0.01	0.05	0.005	0.01	0.05
PSNR (dB)	33.86	33.81	33.40	33.24	33.17	32.71	32.26	32.15	31.68
BER (%)	12.04	12.06	12.36	10.22	10.19	10.37	8	8.02	8.21
NC	0.909	0.909	0.899	0.922	0.921	0.915	0.936	0.936	0.930

TABLE 4.
Quality of watermarked images and retrieved watermark in which different values of the two scaling factors Γ and Δ .

									
Γ	$1.5 \times Q_{u,v}$			$1.7 \times Q_{u,v}$			$2.0 \times Q_{u,v}$		
Δ	0.005	0.01	0.05	0.005	0.01	0.05	0.005	0.01	0.05
PSNR (dB)	34.07	34.05	33.88	32.94	32.92	32.72	31.62	31.57	31.39
BER (%)	20.79	20.84	21.38	18.06	18.13	18.71	15.14	15.14	15.47
NC	0.823	0.821	0.810	0.844	0.843	0.832	0.866	0.866	0.858

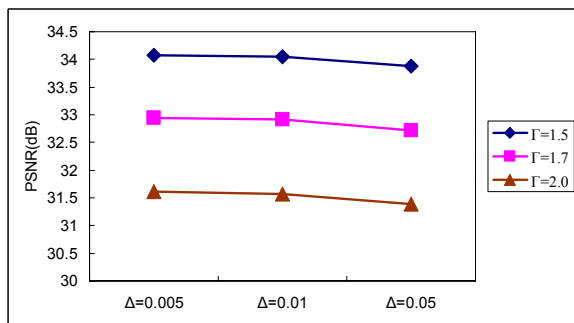


Figure 12. Quality (PSNR) of watermarked images in which different values of the two scaling factors Γ and Δ . (TABLE 4)

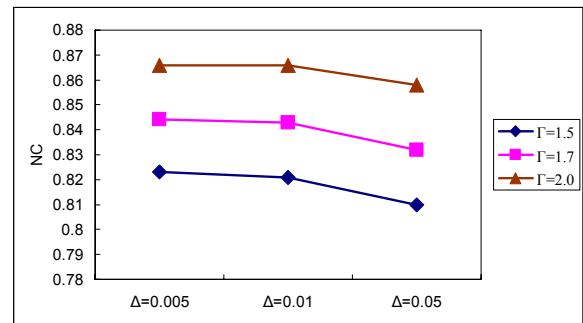


Fig. 14. Quality (NC) of retrieved watermark in which different values of the two scaling factors Γ and Δ . (TABLE 4)

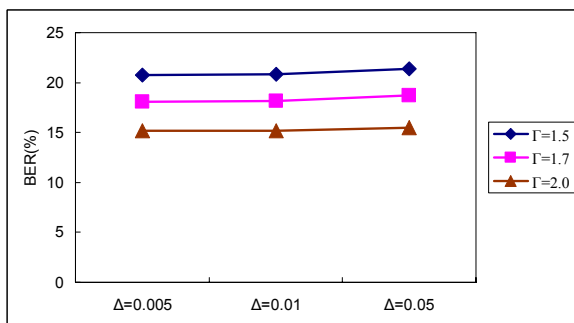






Fig. 13. Quality (BER) of retrieved watermark in which different values of the two scaling factors Γ and Δ . (TABLE 4)

Table 5 shows the quality of the watermarked images and retrieved watermarks, in which different host images. It can be found that the qualities (PSNR) of the watermarked image with respect to the host image are more than 33dB in average for the proposed method.

Now, we describe the results we conducted to analyze the affect of JPEG lossy compression on the watermarked image. The retrieved watermarks extracted from the watermarked images with lossy JPEG compression for the different watermarking techniques are shown. The qualities of the retrieved watermarks under various JPEG compression rate for [8] (resp. the proposed algorithm) are shown in Table 6 (resp. Table 7). It is obviously that

TABLE 5.

Quality of watermarked images and retrieved watermark in which different host image are used for the proposed watermarking technique

The proposed Algorithm		
Watermarked Image		
PSNR (dB)	33.69	33.47
Retrieved Image		
BER (%)	8.98	11.27
NC	0.933	0.909

the proposed watermarking technique can obtain a retrieved watermark with the BER value as low as 17.87% in higher JPEG compression rate 94.9%. The other watermarking technique [8] can give BER value as 22.11% in the same compression rate. Therefore, the proposed scheme is much more robust to resist JPEG lossy compression.

TABLE 6
Quality of retrieved watermark under various JPEG compression ratios for [8]






Watermarked Image				
Compress Rate (%)	91.7	93.8	94.9	95.8
Retrieved Image				
BER (%)	16.22	19.07	22.11	25.86

TABLE 7.

Quality of retrieved watermark under various JPEG compression ratios for the proposed algorithm

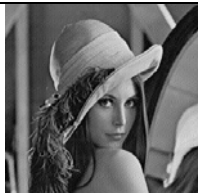




Watermarked Image				
Compress Rate(%)	91.7	93.8	94.9	95.8
Retrieved Image				
BER(%)	9.52	11.57	17.87	25.68

TABLE 8.

Quality of retrieved watermark under attack in rotation for the proposed algorithm






Watermarked Image				
Rotation	90°	180°	270°	360°
PSNR(dB)	32.24	29.69	31.66	33.73
BER(%)	9.89	10.43	9.83	9.48
NC	0.9260	0.9208	0.9269	0.9295
Retrieved Watermark				

Table 9.

Quality of retrieved watermark under attack in resize for the proposed algorithm


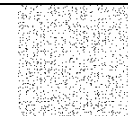
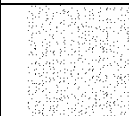

			
Resize	2	4	8
Retrieved Watermark			

Table 8 shows the quality of retrieved watermark under attack in rotation for the proposed algorithm. It can be found that the qualities (NC) of the retrieved watermark with respect to the watermark are more than 0.92 in average for the proposed method.

Table 9 shows the quality of retrieved watermark under attack in resize for the proposed algorithm. It is clearly that the proposed algorithm could not be used to resist the resize attack.

From discuss above, the proposed mechanism for watermarking is robust for the attacks which are JPEG compression and rotation. And, the quality of the watermarked image is still good. In other word, the proposed technique is robust to JPEG compression and rotation. Moreover, the original image and watermark are not necessitated in the extracting process.

ACKNOWLEDGMENT

This work was supported by the National Science Council (NSC), Taiwan, under Contract No. NSC99-2221-E-022-010.

REFERENCES

- [1] B. M. Macq and J. J. Quisquater, "Cryptology for digital TV broad-casting," *Proc. IEEE*, pp. 944-957, 1995.
- [2] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia data-embedding and watermarking technologies," *Proc. IEEE*, pp. 1064-1087, 1998.
- [3] A. Sinha, A. Das, and S. Pandith, "Pattern based robust digital watermarking scheme for images," *Acoustics, Speech, and Signal Processing, 2002 IEEE International Conference on*, pp. 3481-3484, 2002.
- [4] C. T. Hsu and J. L. Wu, "Hidden digital watermarks in images," *IEEE Trans. On Images Processing*, pp. 58-68, 1999.
- [5] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamon, "Decure spread spectrum watermarking for multimedia," *Image Processing, IEEE Trans. on*, pp. 1673-1687, 1997.
- [6] H. Harrak, T. D. Hien, Y. Nagata, and Z. Nakao, "DCT watermarking optimization by genetic programming," *Advances in Soft Computing*, 2006, pp. 347-351.
- [7] C. C. Chang, P. Y. Lin, and J. S. Yeh, "Preserving robustness and removability for digital watermarks using subsampling and difference correlation," *Information Sciences*, 2009, pp. 2283-2293.
- [8] C. M. Kung, J. H. Jeng, and T. K. Troung, "Watermark technique using frequency domain," *Proc. of the 14th IEEE International Conference on Digital Signal Processing (DSP2002)*, July. 2002, pp. 729-731.
- [9] C. M. Kung, J. H. Jeng, and C. H. Kung, "Watermarking Based on Block Property," 16th IPPR Conference on Computer Vision, Graphics and Image Processing, pp. 540-546, Aug. 2003.
- [10] C. M. Kung, K. Y. Juan, Y. C. Tu, and C. H. Kung, "A Robust Watermarking and Image Authentication Technique on Block Property," 2008 International Symposium on Information Science and Engineering, pp.173-177, 2008.
- [11] C. S. Chan and C. C. Chang, "An efficient image authentication method based on Hamming code," *The Journal of the Pattern Recognition Society*, pp. 681-690, 2007.
- [12] Z. M. Lu, J. S. Pan, and S. H. Sun, "VQ-based digital image watermarking method," *IEE Electronic Letters*, pp. 1201-1202, 2000.
- [13] H. C. Wu and C. C. Chang, "A novel digital image watermarking scheme based on the vector quantization technique," *Computers and Security*, pp. 460-471, 2005.
- [14] A. Khan, S. F. Tahir, A. Mahid, and T. S. Choi, "Machine learning based adaptive watermark decoding in view of anticipated attack," *The Journal of the Pattern Recognition Society*, 2008, pp. 2594-2610.
- [15] Z. J. Lee, S. W. Lin, S. F. Su, and C. Y. Lin, "A hybrid watermarking technique applied to digital images," *Applied Soft Computing*, pp. 798-808, 2008.
- [16] T. Todorov, "Improving the watermarking process with usage of block error correcting codes," *Research Report, Institute of Mathematics and Informatics Bulgarian Academy of Sciences*, June, 2005.
- [17] A. Bastug and B. Sankur, "Improving the payload of watermarking channels via LDPC coding," *IEEE Signal Processing Letters*, pp. 90-92, 2004.
- [18] I. Usman and S. Khan, "BCH coding and intelligent watermark embedding: employing both frequency and strength selection," *Applied Soft Computing*, pp. 332-343, 2010.
- [19] H. X. Wang, Z. M. Lu, J. S. Pan, and S. H. Sun, "Robust blind video watermarking with adaptive embedding mechanism," *International Journal of Innovative Computing, Information and Control*, pp. 247-257, 2005.

Huang-Chi Chen received M.S. degrees from Department of Electrical Engineering of I-Shou University, Kaohsiung Country, Taiwan, in 1998. He is currently a Lecturer of Department of Management Information System of Far-East University, and he entered the Ph.D. program in Department of Electrical Engineering of I-Shou University in 2006. His research interests include error correct codes, image process and machine learning.

Yu-Wen Chang received the Ph.D. degrees in electrical engineering from I-Shou University, Kaohsiung County, Taiwan, in year 2006. She joined the faculty of the Department of Electronic Communication Engineering, National Kaohsiung Marine University in year 2007. Her research interests include error correct codes, image process, VLSI, and fuzzy control.

Rey-Chue Hwang received the Ph.D. degree in electrical engineering from Southern Methodist University, Dallas, Texas, U.S.A. in year 1993. He joined the faculty of the Department of Electrical Engineering, I-Shou University at the same year. Dr. Hwang received the outstanding teaching award of ISU in year 2000. He was an IET Fellow and the co-chair of the IEEE CIS Tainan Chapter in 2005-2007. He also served as the member of a council for the Taiwanese Fuzzy System Association in 2005-2006. Currently, he serves as the Chair of the IEEE CIS Tainan Chapter. He is also the director of General Educational Center of I-Shou University now. His research interests include artificial neural networks, intelligent systems, and fuzzy control. He has published over 140 scientific papers in international and national journals or referred conference proceedings.

Image Copy-Move Forgery Detecting Based on Local Invariant Feature

Li Jing, and Chao Shao

School of Computer and Information engineering,
Henan University of Economics and Law, Zhengzhou, China
Email: {jingli776@126.com, sc_flying0527@yahoo.com.cn}

Abstract—Now digital images are widely used in many fields. Making image forgeries with digital media editing tools is very easy, and these image forgeries are undetectable by human eyes. Copy-move forgery is common image tampering where a part of the image is copied and pasted on another parts. Up to now the useful way to detect copy-move forgeries is block matching technique. This paper firstly analyzes and summarizes block matching technique, then introduces a copy-move forgery detecting method based on local invariant feature matching. It locates copied and pasted regions by matching feature points. It detects feature points and extracts local feature using Scale Invariant Transform algorithm. Matching local features is based on k-d tree and Best-Bin-First method. Through analysis we learn computational complexity of the proposed method is similar to existing block-matching methods, but has better locating accuracy. Experiments show that this method can detect copied and pasted regions successively, even when these regions are operated by some process, such as JPEG compression, Gaussian blurring, rotation and scale.

Index Terms—copy-move forgery, block matching technique, local invariant feature, Scale Invariant Transform algorithm, feature matching, k-d tree, Best-Bin-First method

I. INTRODUCTION

With availability of powerful image processing and editing software, it is very easy to tamper digital images and create forgeries which are imperceptible. We have seen many cases about image forgeries in our society, related to news report, academic research, physic, business, law, military affairs and so on. When some image contents are be questioned, people need a tool to distinguish whether the image is real or tampered. In order to satisfy this requirement, various techniques about tampering or forgery detection have been proposed. In the past decade fragile watermarking [1-3] and digital signature [4-6] have become hot research topics to defeat image forgeries. Fragile watermarking requires a watermark to be inserted into the digital image, which needs specially equipped digital cameras and degrades the image quality. And digital signature methods need to generate signature message in advance.

Recently some researchers begin to study a new technology of detecting image forgeries, which is called Passive-Blind Image Forensics (PBIF) [7]. Compared

with fragile watermarking and digital signature technology, PBIF can detect image alteration or identifying the image source without any prior measurement and registration of the image including the availability of the original reference image.

Now there are many ways to tamper image with computer and image editing software, including removal of objects from the image, addition of objects or changing appearance of objects in the image. And the most common is removal of undesired objects from the image, which is done by copying a part of the image itself and pasting it into another part of the same. This alteration is known as “copy-move forgery” [8]. With the purpose of disguising some details, some counterfeiters may perform some post processing on tamper images after copy-move operation, which makes the task of detecting forgery significantly harder. The post image processing includes noise contamination, JPEG compression, blurring, geometrical transformations etc.

A number of methods have been proposed for the detection of copy-move forgery [8, 9, 10, 11]. Theses methods are robust to signal processes such as blurring and Jpeg compression, not robust to geometrical deformation. But the forgery with geometrical deformation is a common attack.

In order to detect the forgery with geometrical deformation we use Scale Invariant Feature Transform (SIFT) algorithm to detect local invariant features of image, then search the matched feature points by SIFT feature matching. If the number of matched points is larger than the assumed threshold value, we would judge the image has been tempered. This method can not only detect copy-move forgery, but also detect copy regions with geometrical deformation, because SIFT feature describes the image local feature better, has great distinctness, and has been proved to be invariant to image scaling and rotation, and also invariant to change in illumination, compression and blur [13].

The rest of the paper is organized as follows. The next section summarizes previously published papers concerned with the topic of this paper. Section 3 reviews and analyzes the SIFT algorithm. Following on from that, the proposed method is explained and each step of the method discussed in detail. Section 5 contains experiments to demonstrate the outcomes of the proposed

method. The last section summarizes the work that has been done in this paper and next research target.

II. RELATED RESEARCH

Copy-move forgery is a very common type of forgeries, which copies some parts of the image and then pasts on another parts of the same image. The purpose of this forgery is to make some objects disappear from the image or forge some objects in image.

The task of copy-move forgery detection is to search the copied regions and their pasted ones. But finding the very same regions is very difficult, because the copied regions usually are processed by some operations such as filtering, noise addition and geometrical distortion. Hence detecting method should detect the duplicated regions, even if they are slightly different from each other.

To accomplish this task several copy-move forgery detecting techniques have been proposed. These techniques can be divided roughly into exhaustive searching and block matching [12]. Exhaustive searching is a direct method, which involves comparison of every image block to other blocks of itself. However, this method would be computationally very expensive and would take $(MN)^2$ steps for an image of size $M \times N$. Furthermore, this type of method might not work in the case where the copied region has undergone some modifications.

Block matching technique is a research focus for copy-move forgery, and many proposed methods belong to this type of technology [8, 9, 10, 11]. In this approach image is divided into overlapping blocks firstly and then detecting the duplicated block pairs, instead of detecting the whole duplicated region, at last use the detected duplicated block pairs to decide the duplicated regions.

Fridrich et al. [8] proposed a block matching detection method based on discrete cosine transform (DCT). It utilized quantized DCT coefficients as the block features, and matched the DCT coefficients of the lexicographically arranged overlapping image blocks. The advantage of DCT is that the signal energy would be concentrated on the first few coefficients, while most other coefficients are negligibly small. Therefore, the changes in high frequencies caused by the operations such as noise addition, compression should not affect these low frequency coefficients.

The method proposed by Popescu et al. [9] is similar to Fridrich's. It used Principal Component Analysis (PCA) instead of DCT to yield a reduced dimension representation. Popescu's method use about half of feature numbers that used in Fridrich's. By doing so, this method is demonstrated to be more effective. But it also has its drawbacks. One of them is that it does not work when the copy region is re-sampled through scaling or rotation.

The both two methods mentioned above divide the original image into blocks and then extract the reduced dimension feature. But sliding window manipulation only shifts a single pixel, and the number of blocks is large.

In order to further reduce computation load, Li et al. proposed a method [10]. This method decomposed the image into four sub-bands using Discrete Wavelet Transform (DWT) firstly, instead of extracting the feature vectors directly from the image blocks. They divided the low frequency sub-band into overlapping blocks to reduce the number of blocks, and speed up the process, based on the fact that most of the energy would be concentrated at this sub-band. They applied singular value decomposition (SVD) on these blocks.

Later Bayram et al proposed to apply Fourier Mellin Transform (FMT) on the image blocks [11]. It first obtained the Fourier transform representation of each block, re-sampled the resulting magnitude values into log-polar coordinates. It obtained a vector representation by projecting log-polar values onto 1-D and used these representations as block features. Their experiments show the method robust to compression up to JPEG quality 20, rotation with 10° and scaling by 10%.

From these methods we can infer that image forgery detection technique based on block matching generally include the following main steps:

- Dividing the image into overlapping blocks,
- Extracting features from each block,
- Searching the similar block pairs,
- Deciding the duplicated regions.

The first step of this technique is dividing the image into overlapping blocks. The second step is a feature extraction process. The biggest challenge of this technique is to determine the features, which would get the same or very similar values for duplicated blocks, even under modifications. Ref. [8] chose quantized DCT coefficients as the block features, in [9], [10], [11] they chose PCA, SVD, FMT coefficients as the block features respectively. These features, except FMT coefficients, are only robust to signal process, not robust to geometrical transformation such as rotation and scaling.

The third step is to find the similar block pairs by feature matching in a reasonable time. In order to save computing time, the references mentioned above used a common match based on lexicographically sorting. The features of each block are inserted into a matrix A , where the rows of the matrix correspond to the blocks and columns of the matrix correspond to the features. If two blocks in the image are similar, their feature vectors therefore corresponding rows in matrix A would be similar as well. If the rows of A matrix is sorted lexicographically these feature vectors would come successively. The corresponding blocks whose feature vectors come successively in the matrix A would be the candidates of block duplicates. The computing time in lexicographically sorting depends on the number and the number of features.

Since the duplicated region would include many overlapping blocks, the last step is to decide the whole duplicated regions. Each block would be moved with the same amount of shift, the distance between each duplicated block pair would be the same as well.

Therefore, the forgery decision can be made only if there are more than a certain number of similar image blocks within the same distance and these blocks are connected to each other so that they form two regions of the same shape.

III. LOCAL INVARIANT FEATURES

From the discussion in section II we learned that determining the feature is very important. Features of the duplicated blocks should be same or very similar, even under some modifications. The robustness of copy-move forgery detection method is based on the invariant of feature. The feature should be invariant to not only signal processes such as blurring and Jpeg compression, but also to geometrical deformation such as rotation and scaling. This paper chooses Scale Invariant Feature Transform (SIFT) feature to be as the matching feature. SIFT feature is a type of local invariant feature, which is proposed by Lowe et al. and proved to be invariant to image scale and rotation, also robust to a substantial range of affine distortion, change in 3D viewpoint, compression, addition of noise, and change in illumination[13].

A. Scale Invariant Feature Transform (SIFT)

SIFT is a method to extract distinctive invariant features from images. It includes key point detector and local feature descriptor.

There are several steps to compute these features as follows: (1) select candidates for features by searching peaks in the scale space from difference-of-Gaussian (DoG) function, (2) locate key points using measures of their stability, (3) assign orientations based on local image gradient directions and (4) calculate the local key point descriptors based on the set of surrounding image gradients. Each feature, a vector with $4 \times 4 \times 8$ elements,

includes coordinates of the detected key point.

B. SIFT Feature used in Copy-Move Forgery

SIFT key point (x, y, σ, θ) represents the key point location in the difference-of-Gaussian scale space, (x, y) denotes the location in image plane, σ denotes the scale in the difference-of-Gaussian scale space and θ denotes the main orientation. Each SIFT key point corresponds a 128-dimensionality feature vector, this paper locating duplicated region by SIFT feature vector matching.

We use SIFT feature because it is invariant to some signal process and geometrical deformation. In addition, some other qualities of SIFT feature also support our choice.

An important aspect of SIFT algorithm is that it generates large numbers of features that densely cover the image over the full range of scales and locations. A typical image of size 500×500 pixels will give about 2000 stable features, although this number depends on both image content and choices for various parameters.

In Fig. 1 (a), (b), (c) and (d) are 4 nature images with size 256×256 pixels, standing for portrait, geometric figure, building and natural scene respectively. (a1), (b1), (c1) and (d1) are SIFT key points extracted from image (a), (b), (c) and (d) respectively, and cross“+” denotes a key points location. In (a1) have 349 SIFT key points, in (b1) have 202 SIFT key points, in (c1) have 585 SIFT key points and in (d1) have 823 SIFT key points.

From Fig. 1 we can find SIFT algorithm extract large number of features. But the features concentrate in the regions with complex structure such as texture areas, corners, and there are few feature points in the smooth area. In this case, if copied region is smooth one, the detecting method based on SIFT features could hardly

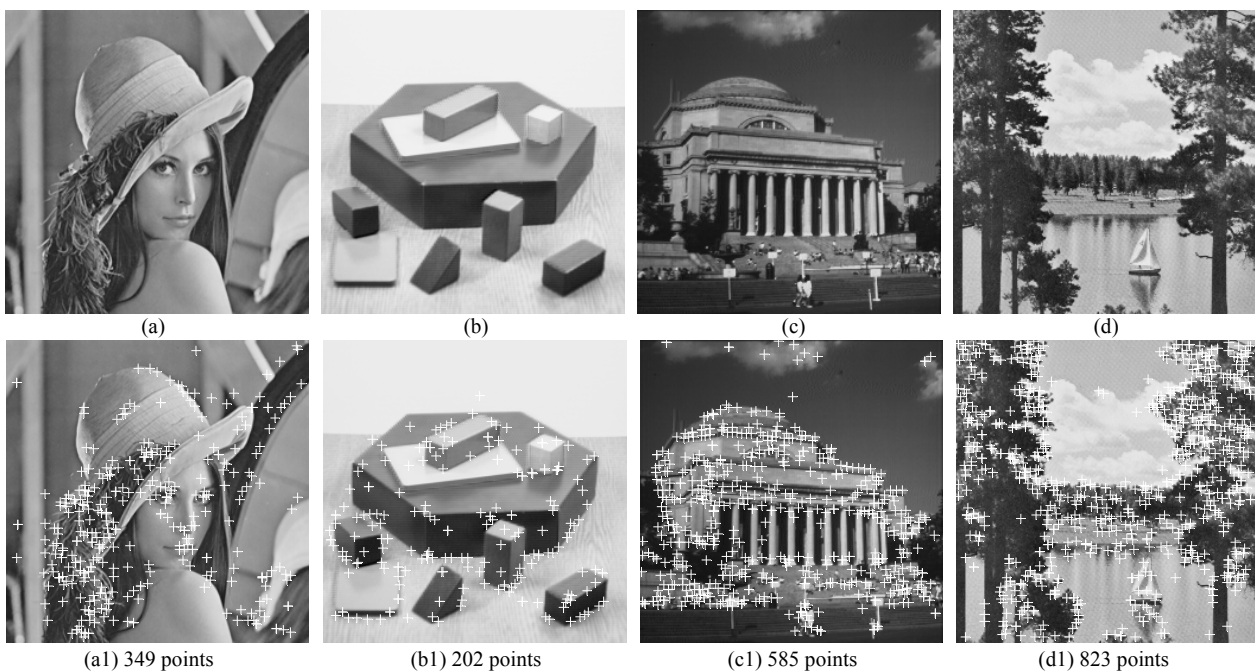


Figure 1. Nature images and their SIFT points

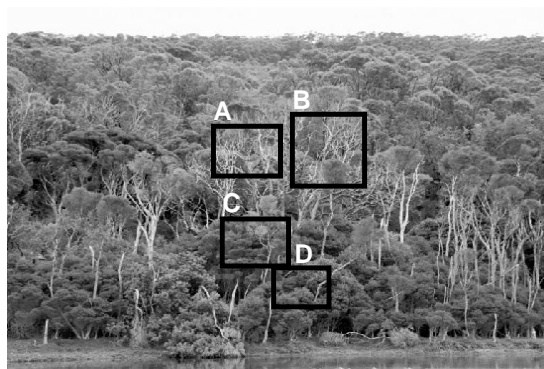
detect the duplicated regions.

In fact for image tamper ideal regions for copy-move forgery are textured areas with irregular patterns, such as grass. Because the copied areas will likely blend with the background, it will be very difficult for the human eye to detect any suspicious artifacts [8].

SIFT features are highly distinctive. Nature sceneries in image, for example trees and grasses, even if are very similar, the SIFT features extracted from them would not same. But if they are the same thing, their SIFT features are same. Even under blur, noise adding, rotation and scale, their SIFT feature would match well. SIFT feature is stable for some modifications.

Fig. 2 is an example showing SIFT feature very distinctive. There are same trees in Fig. 2, and their texture is same. Especially region A and B in Fig. 2(a) have very same structure and texture, same for region C and D. We detect SIFT key points and get corresponding SIFT features in region A ,B, C and D respectively, get 135 key points in region A, 135 key points in region A, 179 key points in region B, 123 key points in region C and 82 key points in region D, as shown in Fig. 2(b). Then match the SIFT features between A and B, C and D respectively, but no find a matched point pair.

Fig. 3 is the example for SIFT feature stability. In Fig. 3(a) the left-top image is copied from the right one, the copied image is rotated 10° and then scaled down 10%. Extract SIFT feature from the two images, and get 319 features from the Left image and 3818 features from right image. Then match features between two image. Fig. 3(b) is the matched result, the number of matched point pairs is 185, and feature matching rate is 60%.



(a)



(b)

Figure2. Example for SIFT feature distinctive.



(a)



(b)

Figure 3. Example for SIFT feature stable.

From the above discussion, we learn the number of SIFT feature is large, but they distribute in areas with complex structure. SIFT feature is distinctive and stable for modifications. Base on these qualities, if finding some matched SIFT feature point pairs in the same nature image, it could judge the image is a copy-move forgery.

IV. PROPOSED DETECTING SCHEME

A. The Proposed Scheme

In section □ we summarized that block matching technique has four main steps, including divide the image into overlapping blocks, extract features from each block, search the similar block pairs and decide the duplicated regions. But our forgery detecting method based on SIFT feature does not need to divide image into blocks. It detects the duplicated region directly, for SIFT feature is local feature, when searching all of matched feature points, the duplicated regions can be decided.

The proposed copy-move forgery detection method has the following main steps:

- Extract local invariant features using SIFT algorithm from image,
- Match these SIFT features each other,
- Create the duplicated region map.

Each step is explained separately in the following.

Step1 Extract SIFT feature

For an input image, if it is an RGB image, convert it into grayscale using the following formula,

$$Y = 0.2989 \times R + 0.587 \times G + 0.114 \times B \quad (1)$$

For SIFT algorithm regards impute image as a matrix, while RGB image has three matrixes. So RGB image need to convert into one matrix to fit SIFT algorithm.

First locate key points use SIFT detector, detected key points group presented with $P = \{P_1, P_2, \dots, P_n\}$. Then generate local feature for every key point use SIFT descriptor, every SIFT feature is 128-dimension vector. Present these feature vectors by $F = \{F_1, F_2, \dots, F_n\}$.

Step 2 Feature matching

In order to detect duplicated regions, every feature vector F_i ($1 \leq i \leq n$) should match with all other feature vectors $\{F_j, 1 \leq j \leq n, j \neq i\}$. Note two matched feature vectors are not exactly same. If their similarity meets certain condition, they can be judged as matched feature.

We use Euclidean distance present the similarity of two feature vectors, denote $\|F_i - F_j\|$. If feature F_i and F_j meet (2), we judge them match.

$$\frac{\|F_i - F_j\|}{\|F_i - F_k\|} < T \quad (2)$$

In (2) F_j is the closest neighbor of F_i , The closest neighbor is defined as the feature vector with minimum Euclidean distance, F_k is the second-closest neighbor of F_i , The second-closest neighbor is defined as the feature vector with second minimum Euclidean distance. T is a threshold, which decide the number of false match. The value of T is more large, more false match will be got.

Identifying the nearest neighbors of feature points in high dimensional spaces is a hard work. The simplest way is exhaustive search, but it is very inefficient and its computational cost is $O(N^2)$.

We use $k-d$ tree to preprocess data into a data structure. $k-d$ tree is a binary tree that stores points of a k -dimensional space in the leaves. If a $k-d$ tree consists of N records, it requires $O(N \log_2 N)$ operations to be constructed [16]

Then use Best-Bin-First (BBF) algorithm to search the closest neighbor. BBF is an approximate algorithm. It can identify the nearest neighbors with high probability using only a limited amount of computation [17].

Step 3 Mark the duplicated regions

The proposed method detect copied regions by SIFT feature matching. Maybe Multi regions are copied and pasted in a same image. In order to make sure the copy-paste region pairs, draw a line between the matched two points. Generally matched points belong to the duplicated region, but false matched points beyond the duplicated region. Matched points densely cover the duplicated regions. In a circular area with radius R centered a matched point, if there no other matched points, we could judge it is false match and abandon this point, on the contrary, mark 3×3 area around the matched point belong to duplicated region.

B. Scheme discussion

Performance of a detection method based on feature matching not only depends on detecting accuracy, computational time is also important factor. Feature extracting and matching take up large computational time of the whole process. Time spent in feature extracting is decided by feature extracting algorithm, and matching time decided by the number of feature.

We compare the proposed method with other homogeneous methods[8,9,10] based on 512×512 image, including feature describe, number of feature vector, feature dimension and locating accuracy. The compared result can be seen in Table I. The feature dimension of our method is 128, exceed that of other method, but its number of feature vector is far less than that of other method, which is the important factor of computational time. And compared with other method, our method has better locating accuracy, which is 3×3 .

V. EXPERIMENT RESULTS

In this section we first show a few examples of copy-move forgery and their corresponding detection result, then give the statistic results of experiment based on a large number of images.

The experiments were carried out with Matlab 7.0. In the feature matching step the threshold T was set to be 0.6, and in the creating duplicated region map step radius R of circular area centered a matched point was set 8 pixels.

A. Some detecting examples for copy-move forgery

These examples include copied and pasted multi-regions and ones with rotation and scale.

The images presented in Fig. 5 were the detection

TABLE I.
COMPARISON OF THE PROPOSED METHOD WITH OTHERS

Methods	Feature describe	Number of feature vector	Feature dimension	Locating accuracy
Fridrich mehod[8]	DCT	$(512-8+1)^2=255025$	64	8×8
Popescu method[9]	PCA	$(512-8+1)^2=255025$	32	8×8
Li method[10]	DWT &SVD	$(256-8+1)^2=62001$	8	16×16
Proposed method	SIFT	About 2000	128	3×3

results of tampered images without any distorted operations. And Fig. 6 presents that of tampered images with rotation and scale. Each row was composed of four images: original image, tampered image, lined copy-move region image and marked tampered image from left to right.

From these experimental results we can see the proposed method can detect copy-move forgery, even tampered region operated by geometrical transform. In Fig. 6(a) test image, the copied region was scaled down by 20% and rotated 30°, then moved another place. The copied region in Fig. 2(b) was scaled down by 30%.

Although the proposed method can detect the forgery

in Fig. 5 and Fig. 6, the numbers of matched feature point pair are very different. In Fig. 5(b), our human eyes could hardly distinguish the tempered regions, but the proposed method gave a better result, detecting many matched feature points. But for the tampered regions in Fig. 6(b), the number of matched points is very small, only 6 matched point pairs. The main reason is that the two images have different texture. The texture structure of the copied region in Fig. 5(b) is very complex while that in Fig. 6(b) is very smooth. SIFT key points distribute the areas with complex texture structure.

B. The statistic results

In order to evaluate the performance of the proposed

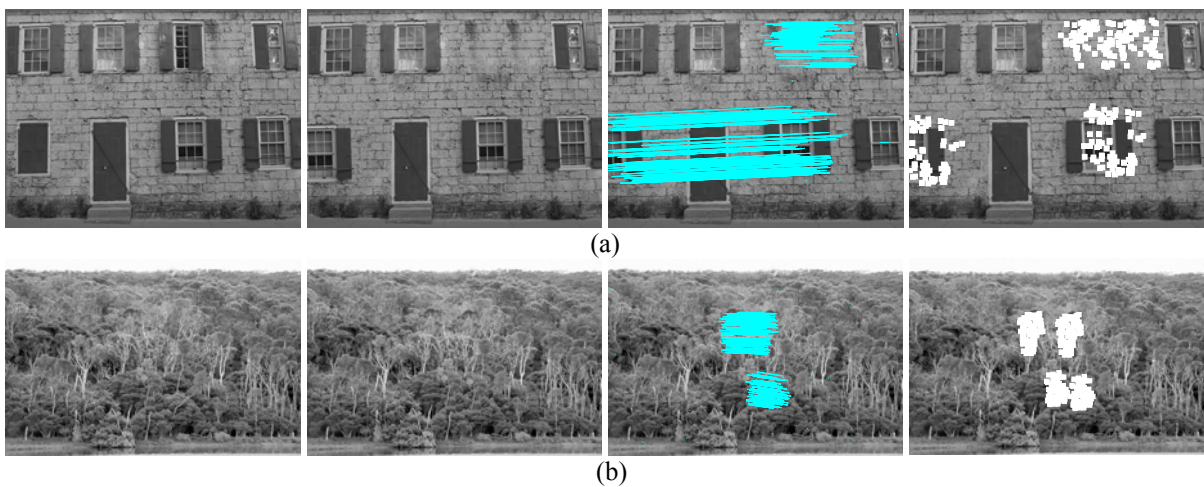


Figure 5. Test images without distort operations and the detection result
(List original image, tampered image, lined copy-move region image and marked tampered image from left to right)

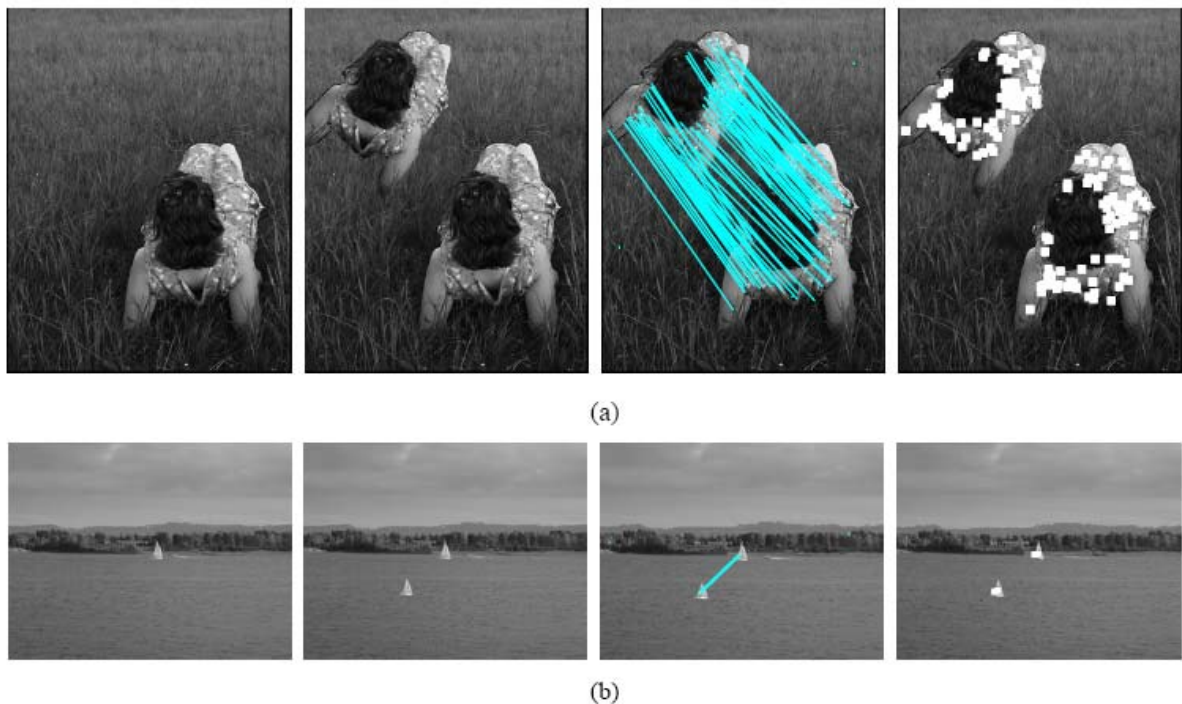


Figure 6. Test images with geometrical distortion and the detection result
(List original image, tampered image, lined copy-move region image and marked tampered image from left to right)

method, we used 200 natural images with size 256×256 to make 200 tampered images by copying a square region at a random location and pasting it into a non-overlapping region. The square region’s size was 32×32, 40×40 and 48×48 respectively. The tampered images were then distorted by different processing operations: JPEG compression with different quality, Gaussian blurring, rotation and scale. If the number of matched feature point pair between two square regions exceeds 6, judge them tamped regions. The experimental results compared with the recent research [16].

In order to learn performance of the proposed method, we quote some evaluation criterions, including accuracy rate, false rate, miss-alarm rate and false-alarm rate. Miss-alarm means the method fails to detect copied or pasted regions, and false-alarm means the method detect some generic regions as copied or pasted ones. They are defined with the following formulas:

$$r_{accuracy} = \frac{|C \cap C'| + |P \cap P'|}{|C| + |P|} \quad (3)$$

$$r_{false} = \frac{|C \cup C'| + |P \cup P'|}{|C| + |P|} - r_{accuracy} \quad (4)$$

$$r_{miss-alarm} = \frac{|\mathfrak{S}_C C \cap C'| + |\mathfrak{S}_P P \cap P'|}{|C| + |P|} \quad (5)$$

$$r_{false-alarm} = \frac{|\mathfrak{S}_C C \cap C'| + |\mathfrak{S}_P P \cap P'|}{|C| + |P|} \quad (6)$$

Where C denotes the set of copied regions, C' denotes the set of detected regions to be copied ones, P denotes the set of pasted regions, P' denotes the set of detected regions to be pasted ones. $\mathfrak{S}_A B$ denotes the complementary set of set B in set A .

To test the performance of our method in the case of detecting JPEG compress image and Gaussian blurring, tampered images were operated by JPEG compression and Gaussian blurring with different parameters. The experiment result was shown in table II. In order to compare with reference [16], table II includes part results of reference [16].

From the statistic data in table II we can see our method has good performance to JPEG compression. The statistic results are not as good as Ref [16] under JPEG compression with Q=80, but it got better statistic data than Ret. [16] under Q=50. And we can see the proposed method has good stability to Gaussian blurring, the accuracy rates $r_{accuracy}$ of three group tests to Gaussian blurring are more than 0.9. That means SIFT feature is stable to JPEG compression and Gaussian blurring. At the same time, we can see the false rates are mainly offered by miss-alarm rates, and nearly all of false-alarm rates equal to zero.

Higher miss-alarm rate means some copied and pasted regions were not detected. And tamped regions not detected by our method are smooth areas. In smooth region SIFT algorithm can not extract more feature, so

TABLE II.
STATISTIC DATA OF THE FORGERY IMAGE OPERATED BY JPEG COMPRESSION AND GAUSSIAN BLURRING

Distorted operation	Evaluation Criterion	32×32		40×40		48×48		
		Proposed method	Ref. [15]	Proposed method	Ref. [15]	Proposed method	Ref. [15]	
JPEG Compression	Q=80	$r_{accuracy}$	0.912	0.961	0.926	0.972	0.925	0.98
		r_{false}	0.147	0.055	0.103	0.04	0.086	0.027
		$r_{false-alarm}$	0.001		0		0	
		$r_{miss-alarm}$	0.146		0.103		0.086	
	Q=50	$r_{accuracy}$	0.813	0.716	0.862	0.714	0.869	0.729
		r_{false}	0.273	0.307	0.231	0.298	0.235	0.277
		$r_{false-alarm}$	0.004		0		0	
		$r_{miss-alarm}$	0.269		0.231		0.233	
Gaussian blurring	$\omega = 3$ $\sigma = 0.5$	$r_{accuracy}$	0.908		0.914		0.928	0.922
		r_{false}	0.172		0.137		0.129	0.091
		$r_{false-alarm}$	0		0		0	
		$r_{miss-alarm}$	0.172		0.137		0.129	

the proposed method based on SIFT can not detected these if these regions are be copied, they could not be detected these tamped regions.

VI. CONCLUSION

Local invariant features are widely used in image recognition and image retrieval. In this paper we introduce them for copy-move image forgery. Since SIFT feature is a type of good local invariant feature with strong stability and distinctness, the proposed method in this paper combines SIFT feature and matching technique based k-d tree and BBF, and gets better performance for copy-move forgery. It can detect tampered regions with some post-operations such as JPEG compression, Gaussian blurring, rotation, scale. But we can see this method fail to detect copied and moved smooth areas, because SIFT algorithm can not extract feature from these areas. Fortunately tampers seldom work in these areas because the forgeries are detected by eyes easily.

Copy-move forgery detection based on local invariant feature is a direct method, its performance is decided by local feature algorithm and matching technology. Next work will concentrated on improving existing feature extracting and matching method.

ACKNOWLEDGMENT

This work was supported in part by a grant from Science Technology Project of Henan, China (No. 0624260019) and National Natural Science Foundation of China (No.092102310163 and 082400410210)

REFERENCES

- [1] MU Celik, G Sharma, and E Saber. Hierarchical watermarking for secure image authentication with localization. *IEEE Transactions on Image Processing*, vol.11, No. 6, 2002, pp. 585-595
- [2] BB Zhu, MD Swanson, and AH Tewfik. When seeing isn't believing. *IEEE Signal Processing Magazine*, vol.21, No. 2, 2004, pp. 40-49
- [3] JH Wu, FZ Lin. Image authentication based on digital watermarking. *Chinese Journal of Computers*, vol.27, No.9, 2004, pp. 1153-1161 (In Chinese)
- [4] GL Friedman. *The trustworthy digital camera: restoring credibility to the photographic image*. IEEE Transactions on Consumer Electronics, 39(4), 1993, pp. 905-910
- [5] CY Lin, SF Chang. A robust image authentication method distinguishing JPEG compression from malicious manipulation. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(2) 2001, pp. 153-168
- [6] CS Lu, HY Liao. Structural digital signature for image authentication: an incidental distortion resistant scheme. *IEEE Transactions on Multimedia*, 5(2) 2003, pp. 161-173
- [7] TT Ng, SF Chang, QB Sun. Blind Detection of Digital Photomontage Using Higher Order Statistics, *Advent Technical Report*, Columbia University, June 2004 (unpublished)
- [8] J. Fridrich, D. Soukal, and J. Lukas. Detection of copy-move forgery in digital images. In *Proceedings of Digital Forensic Research Workshop*, Cleveland, OH, USA, August 2003, pp. 55-61
- [9] A. Popescu and H. Farid. Exposing digital forgeries by detecting duplicated image regions. Technical Report TR2004-515, 2004, Department of Computer Science, Dartmouth College (unpublished)
- [10] Li G H, Wu Q, Tu D. Sun S J. A sorted neighborhood approach for detecting duplicated regions in image forgeries based on DWT and SVD. In: *Proceedings of 2007 IEEE International Conference on Multimedia and Expo*. Beijing, China: IEEE, 2007, pp. 1750-1753
- [11] S. Bayram , ST. Husrev , N. Memon, An efficient and robust method for detecting copy-move forgery, *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 19-24, 2009, pp.1053-1056.
- [12] S. Bayram, H. T. Sencar, and N. Memon, A Survey of Copy-Move Forgery Detection Techniques, *IEEE Western New York Image Processing Workshop*, September 2008, pp. 538-542
- [13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal. of Computer Vision*, 60(2), 2004, pp.91-110.
- [14] R. Sagawa, T. Masuda, and K. Ikeuchi, Effective nearest neighbor search for aligning and merging range images, in: *Proceedings of the 3DIM 2003*, pp. 79-86.
- [15] Beis, J. and D.G. Lowe, Shape indexing using approximate nearest-neighbor search in high dimensional spaces. In *Conference on Computer Vision and Pattern Recognition*, Puerto Rico, 1997, pp. 1000-1006.
- [16] Y. Huang, et al., Improved DCT-based detection of copy-move forgery in images, *Forensic Science International* (2010).

Li Jing, female, born in Henan Province, China on February 20, 1971. She received her PH.D. degree in computer software theory from Information Engineering University, Zhengzhou, China in 2009. Now she is an ASSOCIATE PROFESSOR of School of Computer and Information Engineering, Henan University of Economics and Law, Zhengzhou, Henan Province, China. His research fields include artificial neural network,, data mining and image process, etc. Dr. Jing is also a member of China Computer Federation.

Chao Shao, male, born in Henan Province, China on December 12, 1977. He received his PH.D. degree in computer application technology from Beijing Jiaotong University, Beijing, China in 2006.

He is currently an ASSOCIATE PROFESSOR of School of Computer and Information Engineering, Henan University of Economics and Law, Zhengzhou, Henan Province, China. His research fields include artificial neural network, machine learning, data mining and data visualization, etc.

Dr. Shao is also a member of China Computer Federation.

OWL-S based Service Composition of Three-dimensional Geometry Modeling

Jiangning Yu¹, Hongming Cai¹ and Fenglin Bu¹

School of Software, Shanghai Jiao Tong University, Shanghai, PR China

E-mail: yujne@yahoo.com.cn; hmcai@sjtu.edu.cn; bu-fl@cs.sjtu.edu.cn

Ailing Liu^{1,2}

Shanghai Waigaoqiao Shipbuilding Corporation, Shanghai, PR China

E-mail: y_lal@126.com

Abstract—This paper proposes an OWL-S framework for distributed CAD system based on the combination of semantic web service and CAD technology. Service ontology mapping mechanism is analyzed in detail and semantic model is built with the study of the correlation across the geometry modeling service. On the purpose of accommodating the design pattern of network modeling, this framework supports the service composition of three-dimensional geometry modeling and achieves further integration of service information. At last, a case study from the developed prototype system shows the feasibility and flexibility of this method under distributed CAD environment.

Index Terms—three-dimensional geometry modeling, geometry service, service composition, OWL-S, CAD

I. INTRODUCTION

In the web service based network environment, the development trend of three-dimensional geometry modeling is the integration of CAD technologies and web applications. The integrated system is to realize the packaging and orderly invoking of web service, and improve design efficiency and interoperability. Consequently, the geometry modeling services need to be released freely to provide adequate information for the requirement of uniform interacting mode and integrating heterogeneous systems.

The concept of Web Services for CAD (WSC) aims at interoperation with CAD systems using Web Services [1]. This structure is divided into service Layer, CAD adaptation layer and CAD APIs layer. The service layer provides the access interfaces with WSDL, the adaptation layer packages the XML description of CAD model information for service invoking, and the bottom layer transforms the model data to an identifiable format for CAD system. Parallel WSC [2] is an improved system based on WSC. It uses parallel web services to share assembling information. Through constructing a WSC Master to request WSC node service, it accomplished the assembling of CAD model parts. But it still belongs to a static model information processing method. In the process of modeling, the relationship between model data and structure has been built, so it cannot support the dynamic operations of modeling. Based on this, the

architecture of three-dimensional geometry modeling proposes a mixed solid model for characterization which separate the data manipulation and model display, thus it encapsulates modeling operations into web services. The research of web service based knowledge CAD system [3] presents the method of saving model data in the CAD designing process and parametric modeling with WSIF architecture in order to achieve the goal of knowledge reuse.

According to related researches mentioned above, these structures all use web service as the user interactive mode. They semantically package discrete functions, and provide a loose coupling and reusable graphic modeling interface. After being released on internet, it can be accessed through a standard internet protocol in system. This kind of access will be achieved through a self-contained and self-described application program. Before the semantic web comes up, the distribution of geometry modeling services is discrete and unordered, so the access of service must be based on the one-on-one binding to the local operation. This dramatically decreased the service discovery variance and the service access efficiency. After semantic web has been proposed, the ontology can describe static resources as well as a dynamic operation or service. This enabled users to locate, select, use and compose the web based services. This technology provides more flexibility to the web service based CAD system. By releasing the services to semantic resources, the simple services can be recombined and reused. It will enhance the service based information interaction and knowledge sharing.

II. RELATED WORKS

A. Model Data Exchange

The conventional CAD data formats are designed for single-user mode, such as STEP, IGES, DXF, etc. The system has integrated the geometric information, constraint information and engineering information of the product model. Through parameter, feature or historic modeling, it can simulate product design process. As the model structure is often complicated, the saved model document is usually too large for model rendering and display. And it doesn't have the scene elements either.

Virtual Reality Modeling Language (VRML) [4] is the web language used for virtual scene creation. X3D [5] is an extended technology based on VRML. The main task is to encapsulate VRML function in a light, scalable core. The Web 3D technology creates static 3D models, and adds behavior to these models, which constitute a visual interactive scene [6].

In order to achieve the interaction of CAD data in network environment, there is a need to convert the specific data format to a browser-knowing data format. For example, the conversion from STEP data to VRML is based on the following steps: EXPRESS Language to C++ Language mapping, solid surface triangulation, boundary representation and surface representation, then C++ interface to VRML interface mapping [7].

B. Distributed CAD System

Distributed CAD research is mainly in the field of distributed collaborative design. In this mode, distributed at different locations, product designers and others involving in the network use a variety of computer-aided design tools to conduct collaborative activities. Each user in the activities can feel the presence of other users with different levels of interaction. Research in this direction experienced network communication, distributed computing and computer supported cooperative work. After a simple process of combining CAX/DFX technology [8] and web technology, in recent years, it turns to some deep, core technology issues. Service-oriented CAD system architecture asks user to connect CAD system applications through public web service interfaces. With service invocation at the clients, it achieves the interoperation of visualization application in the internet environment.

C. Semantic Technology

Semantic web service mainly contains three technology subjects: Ontology Web Language for Services (OWL-S) [9], Web Service Modeling Ontology (WSMO) [10] and WSDL-S [11]. A comparison is shown in Table I.

TABLE I. A SIMPLE COMPARISON

Item	OWL-S	WSMO	WSDL-S
Service	General	Web Service	
Semantic	From Semantic Service Description to Web Service		Web Service to Semantics
Interaction Model	Description Logic	Framework Logic	No Model

OWL-S and WSMO are similar in terms of semantic matching and information extraction, but the interactive mode is different. OWL-S is already a W3C recommendation. It is mainly a logic description using process model. WSMO is mainly a framework description using a mediator meaning that WSMO service ontology is more suitable for the concept modeling. WSDL-S is an extension of present web service standard, you can use different semantic language to mark the web service, but still comply with WSDL specification. So the

expansion is limited. That way if you need to use the original service in a composition mode, new rewritten WSDLs must be provided [12][13][14].

For the layer of service implementation, OWL-S provides rich semantic interfaces. OWL-S FLOWS define the combination rules for the web service [15]. In addition, OWL-S does not care about the specific service type. The current service matching engine is established based on WSDL. After replacing the release standard of service, the loose coupling logical framework still can be used. Comparing with the other semantic technology, it is more suitable for the service composition application of three-dimensional geometry modeling.

III. OWL-S FOR CAD

OWL-S for CAD (OSC) provides a solution for OWL-S based CAD system. The ontology framework of OWL [11] provides a people-oriented understanding of the semantics and this creates the framework of the description to be customized and shared. Various Web CAD site should use a basic set of classes and attributes for the statement and description of ontology. Thus it calls local CAD engine interface to respond to the requested ontology service which integrates web service resources in accordance with service process by OWL-S Process technology. OWL-S provides rich semantic interfaces which can both describe a simple web service, and a complex combination of multiple services [16].

For composite services there should be an interactive session between the user and provided services so that the user can make a choice and provide conditional information. Modeling services use this information and save model data into VRML coordinate index node in the form of XML. The VRML model is used to be transferred into the X3D scene rendering model by VRML to X3D Translation [17] which can bring a user-friendly internet environment for the interaction of geometry scene data.

In the prior research, most of them focused on the web service based encapsulation of CAD interface. OSC uses web service ontology described by owl to establish ontology service model. Basic framework as shown below: The same as WSC framework, the under layer is service encapsulation layer. But in OSC framework, the packed services should be fundamental and static so that they can be used for the upper ontology service mapping and service composition. From the CAD design perspective, model elements define the data and conditions during service invocation. Service correlation determines service composition approach, which is the integration from a lower level to the top.

An OWL-S service [18] is described by: ①Service Profile describes the information of service organization, functions and characteristics of properties (Service Parameter, Service Category), etc. ②Service Model describes the work nature of geometry modeling service as a process. The main entity type of process ontology is the process. ③Service Grounding is the mapping from the abstract definition to the concrete realization of the service description which specifies the details of how to

access services, including protocols and message formats, serialization and positioning. The owl-s service is a further abstract of design process and will be reflected to the effective service hierarchy. As the original atomic service being combined, it ultimately provides more advanced modeling services for the network environment. The following figure shows the OWL-S for CAD framework. The left side is geometry modeling service hierarchy and the right side is service encapsulation architecture.

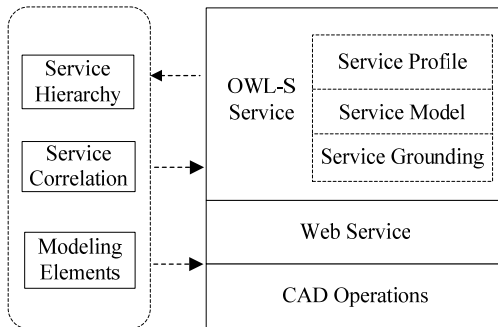


Figure 1. OWL-S for CAD Framework

IV. GEOMETRY MODELING SERVICE

A. Modeling Elements

On the basis of overall design, the task of three-modeling is to draw the specific model structure according to the established principles to reflect the functions to be achieved. In general, the abstract working principle into a certain model needs to determine material, shape, dimensions, tolerances, and disposal method of the structure. Among them, the material, size, tolerance of the components is model data. The shape of components is model structure. The processing mode is operating condition.

(1) Model Data

To build a model you need to know the basic data including absolute position, relative positions, dimensions, tolerances. (Used only as identification, not data processing)

(2) Model Structure

Model structure is mainly for the expression of three-dimensional geometry model and the physical shape characterization of structure and the correlations between entities. The correlations include the direct and indirect correlations. The solids contacted with the surface have the relation of direct correlation. For example, gear and shaft have a consistent centerline with the inner surface and the outer surface fitting together. The indirect correlations include location correlation and motion correlation. Location correlation indicates the relative position of entities, such as the car position relationship between two parallel axles. Motion correlation indicates the probable movement locus in the scene. For example the movement locus of cutting tool should be parallel to the centerline of the principal spindle which illustrates the indirect correlation.

(3) Condition

Condition is also crucial for the modeling process. A model with full information not only need the model data and structure, but also need to identify the modeling constraints and rules, such as the scene rendering and illumination conditions.

Therefore, from the service encapsulation perspective, geometry modeling services should include the operations of the three modeling elements which provide the input and output of model data, form the basic model structure and build the scene conditions in modeling process. The encapsulated operations should contain the data, structure and condition elements.

B. Five-Stage Modeling and Service Correlation

CAD design guidelines require the analysis of product functionality to get the product model design. The modeling process of a product is subject to five-stage design process.

Stage 1: Make the analysis of the basic functions of the product to determine the structure and the surface features of the shapes.

Stage 2: Determine the physical location and the connection relationship of the shapes.

Stage 3: Link the body surface and other surfaces together to form the basic parts of a product.

Stage 4: Assemble the product parts to each other to form the main components.

Stage 5: Determine the space constraints and the dimension conditions of installed products components

Through the analysis of three-dimensional geometry modeling elements and modeling process, we have the five stages of modeling service, which are defined for different service layers from S1 to S5.

S1: The service sets SKETCH, CURVED SURFACES, PRIMITIVES accept the inputted model data and generate entities. Although here a sketch model is two-dimensional graph, but its type in the system is still the type of entity which can be used for other entity operations. The output of surface modeling and primitive modeling is an entity type.

S2: The service set FEATURE creates the features of the model. The service set TOPOLOGY operates on the entities through the Boolean operator and form the link between different surfaces to create a product part.

S3: The service set TRANSFORMING repositions the model structure by coordinate transformation, mirror or rotation.

S4: The service set ASSEMBLING assembles the parts through their matching relationship to form components.

S5: The service set SCENE RENDERING adds light, shadows, transparency effects and other conditions to the product model components.

These services in accordance with relationship of the entities can provide a complete service system as shown in Fig 2.

Following the procedure idea of CAD design procedure, modeling operations should have parameters with uniform type. That is the type of a model which is defined as the solid type. First the CAD meta-model is

constructed by the sketches, the basic entities and the surface modeling set. Then the meta-model data as actual parameter and the model type as formal parameter can be passed to the topological operation, feature operation and transform operation to construct a solid model that is used for the final assembling and Scene.

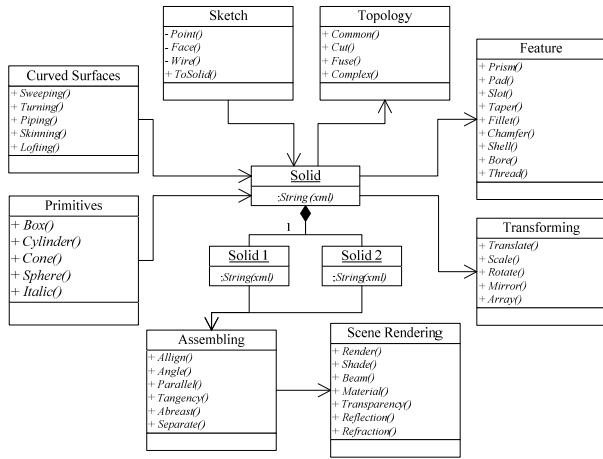


Figure 2. Service Correlation

C. Service Hierarchy

Making analysis of modeling elements and ascertaining the layers of service system will be useful. It contributes to network visualization environment to help providing common and standard functional modeling services that will greatly enhance the efficiency and consistency of the distributed design environment. However, due to different design tasks, companies may generate more advanced modeling services based on the use of common services. Different service parameters and conditions will allow service providers to provide many services with similar features but different parameters. Complex modeling using basic services repeatedly will be bound to result in information redundancy and waste of resources. Therefore, design companies combine the atomic services to provide a higher level of functional services to network environment. It will further optimize the network service resources and improve the design efficiency. For example, a car model designer may use different parts service to complete the wheel modeling. However, these parts service can be combined as the wheel part service for the vehicle designers to use. They no longer need a decomposition of the design task and use those lower layer services. The following diagrams represent the optimization result of the composition of five-layer service system.

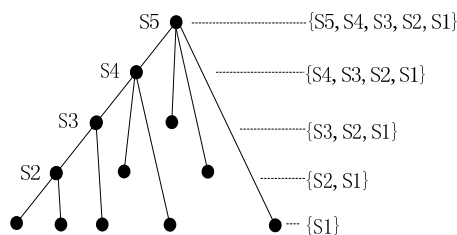


Figure 3. Service Hierarchy without Composition

As shown above, without using the composition of service, S2 design tasks need to use the {S2, S1} set of services, S3 design tasks need to use {S3, S2, S1} set of services and so on. The more high-level the services are, the higher the complexity is. S5 design task need to use all of the lower layer services. The purpose of service composition is to combine lower layer services in accordance with the designing task forming a new service for the invocation by the high-level service. It will reduce inter-service invocation between different service levels. As shown in Fig 4.

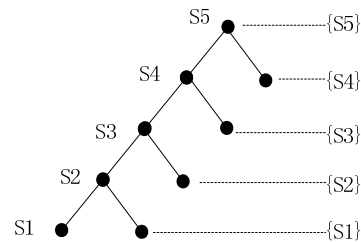


Figure 4. Service Hierarchy with Composition

Higher layer services make the composition of lower layer services. Functionally the combined service has an inclusion relationship with the lower layer service. The design tasks at each level only use the peer service, not access to the cross-layer service, thereby reducing the frequency of using underlying service, while improving the functional requirements of design process. Under the premise that without affecting the complexity of design tasks, the underlying service will be more standardized, and the functionality of advanced service will be more centralized.

V. ENACTMENT

A. Ontology Service

Ontology model of three-dimensional geometry services includes service description, product description, input, output, preconditions, effects, access conditions, service quality, and safety parameters and so on. In OWL-S, a service is described by Service Profile, Service Model and Service Grounding [19]. The service profile presents the basic information of ontology service. The service model describes service definitions. The service grounding supports the ontology service.

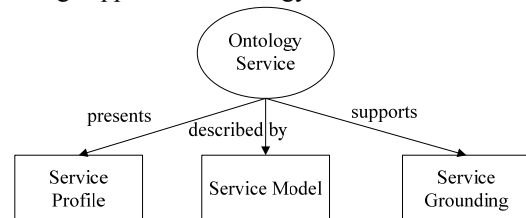


Figure 5. Ontology Service Framework

(1) Service Profile

First, it describes the basic information of service provider including serviceName, textDescription and contractInformation. The service provider is mainly CAD design companies.

Second, it describes the function of service. Mainly contains IOPE: Inputs, Outputs, Preconditions, and Effects. Inputs, Outputs are the data flow of service. Preconditions, Effects are the status flow of service. OWL-S can define the conditional expression of Outputs and Effects. Until some conditions are satisfied, the Outputs and Effects could be generated.

Third, it can describe the category and the QOS information of service. Service profile defines the performance attribute by serviceParameter, describes the classification by serviceCategory.

(2) Service Model

Implement the CAD design process based on web service can be seen as an ordered invocation of geometry modeling services. As these functional services are combined and mapping into ontology, the ontology service is produced. It is seen as the process. Service model describes the internal processes of these services.

It is the rules of OWL-S. The operated object of modeling service is model entity. The data of model structure is saved as the VRML nodes in XML. In the procedure of service interaction, the XML document is parameter and its type in owl is string.

In service model, Processes are connected to the IOPEs in service profile. The object case of IOPEs is created in Service Model, through these definitions: hasParameter, hasInput, hasOutput, hasPrecondition, hasEffect [20]. Here is an example of the IOPEs definition of Topology Service showing in Table II.

TABLE II. TOPOLOGY SERVICE IOPEs

Topology Service : BooleanCut	
Parameters	Inputs, Outputs
Inputs	Shape1, Shape2
Outputs	Shape3
Precondition	Shape1 and Shape2 intersect
Effect	Shape1 is cut by Shape2

OWL-S's Process is divided into three categories: Atomic Process, Simple Process and Composite Process [21]. An atomic process is a description of a service that expects one message and returns one message in response. A composite process is one that maintains some state, each message the client sends advances it through the process. Simple process is an abstract concept. An atomic can realize a simple process, and a composite process can collapse to a simple process. Atomic process is the process that cannot be divided. Each atomic process has grounding information that describes how to access the process. Composite Process consists of a number of atomic processes and service processes. OWL-S's ControlConstruct defines the implementation order of each sub-process. The ControlConstruct are Sequence, Split, Split+Join, Unordered, Choice, If-Then-Else, Iterate, Repeat-Util and so on [22].

In terms of CAD design, when the service is request, an operation is triggered. So we will rarely use choice or if-then-else control construct. On the basis of five-stage modeling service, the composite process occurred inside one stage should use split or other unconditional

construct. But during cross-stage composition, the composite process must use the sequence construct. That is caused by the service hierarchy, as advanced functions are generated on the basis of fundamental functions.

(3) Service Grounding

The grounding of a service specifies the details of how to access the service. Service grounding refers to the service access protocol, message format, port and so on. It packed abstract description of input and output in the atomic process into a network message. The grounding concept is similar as binding concept in WSDL. OWL-S uses WSDL to describe the specific information, so it is necessary to implement concept mapping between OWL-S and WSDL.

B. Service ontology mapping

Service ontology mapping includes two aspects. On one hand, the description of modeling service using WSDL is syntax description. It describes the input and output but doesn't describe the service preconditions. In the result, the discovery and matching of the modeling services require human intervention. However, using OWL-S can add semantic description to the original modeling service, in particular, the grounding description will let computer automatically find and match the corresponding service. On the other hand, the OWL process can describe logic relations in the service access process, which means with OWL logic syntax fine-grained sub-services can be combined and optimized into coarse-grained service resources for a re-release. Consequently there are more and more provided services, but the use of them will be increasingly simple because service functions are integrated.

According to W3C specifications, mapping WSDL to OWL [23] should keep to certain standards. Therefore, service ontology mapping of three-dimensional geometry modeling should satisfy specific conditions. As shown in Table V. Specifically, the atomic process corresponds to modeling service operations. Both inputs and outputs of OWL-S atomic process correspond to modeling service messages. The type of inputs and outputs (OWL class) corresponds to the abstract type in WSDL that is taking WSDL as a class of specific operation definitions including definitions of parameters and types using in modeling service. OWL uses RDF/XML to encode abstract service. With the mapping specifics and logical relationship, a number of WSDL services are packaged into an OWL resource to be re-released.

TABLE III. SERVICE ONTOLOGY MAPPING RULES

WSDL	OWL
operation	atomic process(one-to-many)
message body	inputs/outputs set
message part(abstract type)	owl class
wSDL binding	owl grounding
xml schema	description logic

C. Service composition

The development trend of semantic service is simplification and automation. The basic unit of service composition is the scattered atomic service. OWL-S ontology model contains four parts: service name, service description, service provider, and service URL. The mapping from web service to ontology service and composition need the definition of the process. It converts old services to new descriptions, validates and caching ontology, then executes service process. For the purpose of using OSC for ontology service composition to construct geometry model on internet, the system need to define the interface for the service requestor to access. OWL-S APIs [24] has two important interfaces, OWL Ontology and OWL Knowledgebase. OWL Ontology represents the stored information in a single file. The model of RDF data can be loaded on OWL Ontology. Only ontology object can be used for composition.

TABLE IV. SERVICE COMPOSITION PSEUDO-CODE

Service Composition	Service Execution
<pre> Create a CADService Sequence(List, URI){ Create an Ontology; Create a CADService from URI; Create a Composite Process(URI); Create a Sequence Construct; Put the Sequence into Composite Process; Foreach (Size of Services){ Get CADService from the List; Get Process from the CADService; Create a Perform Construct; Put the Process into Sequence;} Create Profile; Create Grounding; Return CADService; } </pre>	<pre> Class RunCADService{ Dim CADService; Dim CADOperation; RunCADService(){ Create an OWLKnowledgeBase; Read a CADService from the URI of Owl; Get the Process of the CADService; Create an empty ValueMap; Set the input Parameters; Create an Execution Engine; Get the Result; System.out.println(Result); } } </pre>

VI. PROTOTYPE SYSTEM AND EXPERIMENT

A. Prototype system

The following figure shows the architecture of the prototype system.

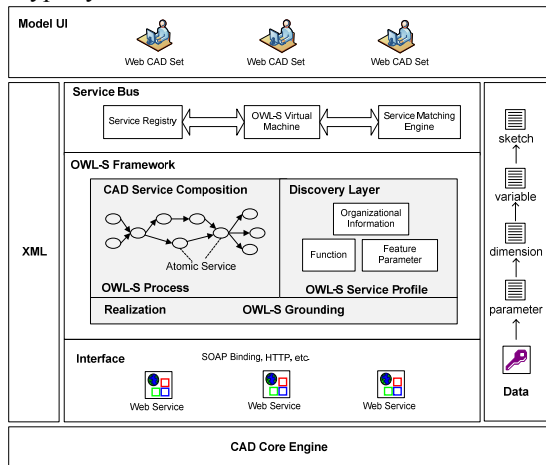


Figure 6. System Architecture based-on OWL-S

The overall framework for the system is Client/Server architecture including three parts: the application program interface, the client and the server-side.

System execution engine is based on CAD core engine which provides the basic interface for graphic modeling. The interface will be used for the encapsulation of web service. Server-side is based on OWL-S framework. At the beginning encapsulated services are unordered in the network environment, so the service's process model and the semantic logic input and output should be matched through the description mapping from WSDL into OWL. Then services are deployed in the bus of server-side which connects service registry and service matching engine with OWL-S virtual machine [25]. The client is distributed web CAD set which customs the matched service to generate and view the design models.

The prototype system uses CAD engine interfaces to implement the encapsulation of modeling operations and releases services with service publishing tool (Axis). Each web service will be deployed on the server-side, and be accessed by internet browser with the mode of SOAP binding, HTTP, etc. At the client, WSDL documents are mapping into corresponding owl document, execution process is defined with OWL process sequence mode combining atomic process services. The output result is saved in VRML nodes as string type. The core of user system uses VRML to X3D exchanger to render VRML nodes into X3D scene model. The prototype system UI shows as bellow.

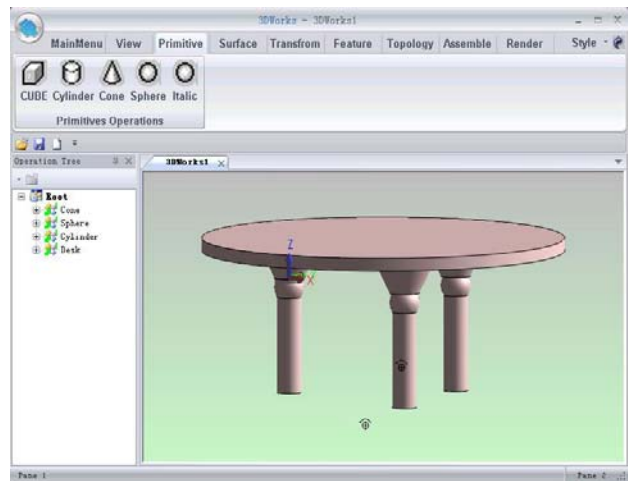


Figure 7 UI of the Prototype System

B. Case study

Take a 3D desk modeling process as an example. Use Opencascade6.3 as the CAD engine. The basic implementation of modeling operations is through Dynamic Link Library (DLL). The process mainly consists of four steps.

- (1) Create leg primitives. Use Cone, Cylinder, Sphere in Primitives set.
- (2) Fuse primitives. Use Fuse operation in Topology set.
- (3) Copy leg one and translate. Use Copy, Transform operations in Transforming set.

(4) Finish desk modeling.

The geometry modeling services describe these different kinds of activities in the design case. We start with creating instances of the atomic process. Define the used web services as atomic processes. The decision what is a composite process can be further refined into a combination of atomic processes. Therefore decide what are the inputs and outputs for each of the atomic processes. The following diagram is an example of service publishing for the atomic process which defines the input and output parameters.

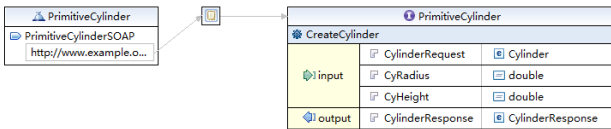


Figure 8. an Example of Service Publishing

Service model is the core for OWL-S based service. Along with start of the process, the atomic processes PrimitiveCone, PrimitiveCylinder and PrimitiveSphere make up a composite process. The ControlConstruct signs before and after are Split and Split+Join. These service processes belong to the S1 layer in five-stage service system. The control construct sign between the TopologyFuse process and Transforming process is Sequence. The two service processes belong to the S2 layer and S3 layer respectively.

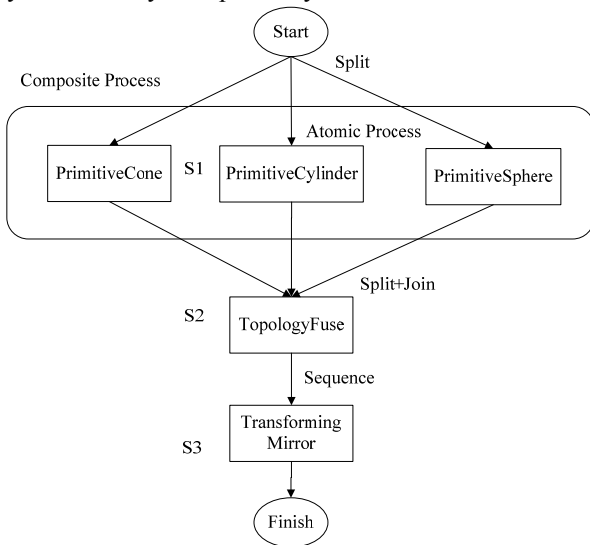


Figure 9. Service Model

Based on the definition of composite process, ontology mapping and service composition of the three-dimensional geometry modeling is a transition process from the old description into the new description. The owl-s engine will verify and buffer the ontology object to execute the process. As shown in Fig.8, the PrimitiveCylinder service is written in WSDL. For the semantic service composition it is rewritten in OWL based on the mapping rules. A fragment is shown as following. It describes the service ID: CylinderService, the service atomic process: CylinderProcess.

```

<?xml version="1.0"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns="http://www.example.org/owl/Cylinder.owl#"
  xmlns:expression="http://www.daml.org/services/owl-s/1.2/generic/Expression.owl#"
  xmlns:service="http://www.daml.org/services/owl-s/1.2/Service.owl#"
  xmlns:process="http://www.daml.org/services/owl-s/1.2/Process.owl#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:swrl="http://www.w3.org/2003/11/swrl#"
  xmlns:grounding="http://www.daml.org/services/owl-s/1.2/Grounding.owl#"
  xmlns:daml="http://www.daml.org/2001/03/daml+oil#"
  xmlns:profile="http://www.daml.org/services/owl-s/1.2/Profile.owl#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:list="http://www.daml.org/services/owl-s/1.2/generic/ObjectList.owl#"
  xml:base="http://www.example.org/owl/Cylinder.owl">
  <owl:Ontology rdf:about="">
    <owl:imports rdf:resource="http://www.daml.org/services/owl-s/1.2/Process.owl"/>
    <owl:imports rdf:resource="http://www.daml.org/services/owl-s/1.2/Service.owl"/>
    <owl:imports rdf:resource="http://www.daml.org/services/owl-s/1.2/Grounding.owl"/>
    <owl:imports rdf:resource="http://www.daml.org/services/owl-s/1.2/Profile.owl"/>
  </owl:Ontology>
  <service:Service rdf:ID="CylinderService">
    <service:presents>
      <profile:Profile rdf:ID="CylinderProfile"/>
    </service:presents>
    <service:describedBy>
      <process:AtomicProcess rdf:ID="CylinderProcess"/>
    </service:describedBy>
    <service:supports>
      <grounding:WellGrounding rdf:ID="CylinderGrounding"/>
    </service:supports>
  </service:Service>
  <profile:Profile rdf:about="#CylinderProfile">
    <service:presentedBy rdf:resource="#CylinderService"/>
  </profile:Profile>
  </service:Service>
  </owl:RDF>
  
```

Figure 10. an OWL Fragment

In the above case, the use of OWL-S to implement a distributed CAD system, not only achieve the main mechanism of Web Service for CAD, but also has advantages of the semantic expansion and service composition. The prototype system verifies the feasibility and flexibility of the framework, encouraging the development of distributed CAD system under semantic web environment. However, present study mainly focused on the realization of applications in the service composition. Along with the development of semantic web technology, further research can focus on the discovery and matching of geometry modeling services, semantic constraint model and the test method to perfect the system functionality.

VII. CONCLUSION

Semantic web service, built on the open standard, introduces a new computing concept by enabling different clients to access their services. Users and software agents should be able to discover, invoke, compose, and monitor Web resources offering particular services and having particular properties, and should be able to do so with a high degree of automation if desired. According to the current development situation of distributed CAD systems under internet environment, this paper provides a solution to three-dimensional geometry modeling from the semantic web aspect. It establishes an OWL-S based service composition framework and defines the service model. By studying the mechanism of mapping modeling service into semantic ontology, it proposes service composition application of the ontology resources in a procedure way, thereby the problems of scattered resources and service integration difficulties existing in current service-oriented distributed CAD systems can be solved effectively. In the end, a practical example and the prototype system demonstrate the implementation process of composition application of OWL-S based modeling services.

Further study should be focus on the following point:

(1) Adding semantic annotation to the geometry modeling services to assist service discovery and service matching.

(2) For the automatic ontology service discovery and invocation, there is need to create a series of upper ontology.

(3) For the scene modeling and model data interaction, service work flow and data base support are needed.

Although the current prototype system did not address these issues, further work should focus on them.

ACKNOWLEDGMENT

This paper is supported by the National High Technology Research and Development Program of China ("863" Program) under No.2008AA04Z126, the National Natural Science Foundation of China under Grant No.61073086 No.60603080, and Shanghai Science and Technology Projects 09DZ1121500.

REFERENCES

- [1] Byungchul Kim and Soonhung Han, "Sharing of CAD assembly model data using parallel Web Services", 12th International Conference on Computer Supported Cooperative Work in Design, 2008, Page(s): 434 – 440.
- [2] B. Kim and S.Han, "Retrieval of CAD Model Information Using Web Services", Proceedings of Design Engineering Workshop 2007, RCAST, the University of Tokyo, Tokyo, Japan, July 26-27, 2007, pp. 26-29.
- [3] Jiangning.Yu, "Knowledge-based Web Service Environment Architecture for Network 3D Visualization", Advanced Materials Research Vols. 102-194 (2010) pp 926-930.
- [4] "VRML Virtual Reality Modeling Language", - <http://www.w3.org/MarkUp/VRML/>.
- [5] X3D Development, <http://www.web3d.org/x3d/>.
- [6] Web 3D, <http://www.web3d.org/>.
- [7] Sun Hongwei, Wang Jian, "a Fast Algorithm of Entity Triangulation in Format Transforming from STEP to VRML", Mechanical science and technology, 2001, 20(4).
- [8] A. Khaled, Y. Ma, J. Miller, A Service Oriented Architecture for CAX Concurrent Collaboration, 4th IEEE Conference on Automation Science and Engineering, Key Bridge Marriott, Washington DC, USA, August 23-26, 2008.
- [9] "OWL-S: Semantic Markup for Web Services", <http://www.daml.org/services/owl-s/1.0/>.
- [10] W3C Web Ontology Language, <http://www.w3.org/tr/owl/features>.
- [11] "Web Service Modeling Ontology (WSMO)", <http://www.w3.org/Submission/WSMO/>.
- [12] "Web Service Semantics - WSDL-S", <http://www.w3.org/Submission/WSDL-S/>.
- [13] W3C Web Services Description Language, <http://www.w3.org/TR/WSDL/>.
- [14] Web Service Modeling Ontology, <http://www.wsmo.org/>.
- [15] Web Ontology Language for Web Services, <http://www.daml.org/services/owl-s>.
- [16] OWL Web Ontology Language, <http://www.w3.org/TR/owl-features/>.
- [17] VRML to X3d Translation, http://ovrt.nist.gov/v2_x3d.html.
- [18] J.Sciocluna, C.Abela and M.Montebello, "Visual Modeling of OWL-S Services", CSAW, Computer Science Annual Workshop, September 23-24, 2004, pp.92-100.
- [19] Massimo Paolucci, Katia Sycara, "Autonomous semantic web services", IEEE INTERNET COMPUTING, 2003.
- [20] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Groszof, and M. Dean. Swrl: A semantic web rule language combining owl and ruleml, 2003. Available at <http://www.daml.org/2003/11/swrl/>.
- [21] Aaron Swartz, "MusicBrainz: a semantic Web service", IEEE INTELLIGENT SYSTEMS, 2002.
- [22] S.McIlraith, T.C. Son, and H.Zeng. Mobilizing the Web with DAML-Enabled Web Service. In Proc. Second Int'l Workshop Semantic Web (SemWeb'2001), 2001.
- [23] D.Martin, M. Burstein, O. Lassila, M. Paolucci, T. Payne, and S. McIlraith. Describing Web Services using OWL-S and WSDL. <http://www.daml.org/services/owl-s/1.1/owl-s-wsdl.html>, 2004.
- [24] E Sirin, OWL-S API, <http://www.mindswap.org/2004/owl-s/api/>
- [25] M. Paolucci, A. Ankolekar, N. Srinivasan, K. Sycara and P.F. Patel-Schneider, "The DAML-S virtual machine", International Semantic Web Conference. Volume 2870 of Lecture Notes in Computer Science, Springer (2003) 290–305.

Jiangning Yu was born in 1986; he received his BS degree in School of Mechanical Engineering in 2008 from Shanghai Jiao Tong University, China. He will receive MS degree in School of Software in 2011 from Shanghai Jiao Tong University, China.

He won the 1st prize of National Robot Contest in 2006, China. He published the paper "Knowledge-based web service environment for 3D visualization" in Journal of Advanced Materials Research (EI indexed) in 2010. In the same year, he got the award, "Excellent Teaching Assistant", for the course of Modeling of Enterprise and Information System.

He is currently studying in Information System Technology Laboratory as a postgraduate student in Shanghai Jiao Tong University, China. His research interests include Geometry Modeling Technology and Service Oriented Architecture.

Hongming Cai was born in 1975, he received his BS degree in Aircraft Structural Strength Design in 1996, and received his MS degree in Machinery manufacturing automation in 1999 and Ph.D. degree in Aerospace Manufacturing Engineering in 2002 from Northwestern Poly-industrial University, China. His research interests include Computer Aided Design and Computer Graphic, Information Integrated Technology. He served as postdoctoral research fellow at the Computer Science and Technology Department of the Shanghai Jiao Tong University, China during the period of November 2002 to September 2004. And he served as visiting professor at the Business Information Technology Institute of University Mannheim, Germany during the period of August 2008 to July 2009. The visiting scholarship was appointed and sponsored by Alfried Krupp von Bohlen und Halbach Foundation, Germany. Dr. Cai is currently an associate professor in School of Software, Shanghai Jiao Tong University. Dr. Cai is a director of China Engineering Graphics Society, and he is a member of ACM and a senior member of China Computer Federation.

Fenglin Bu was born in 1961; he received his BS degree in School of Material Engineering in 1982 from Shanghai Jiao Tong University, China. He received his MS degree in School of Material Engineering in 1995 from Shanghai Jiao Tong University, China. He got the 3rd prize for Process in Science and Technology of State Education Commission in 1998, China. He got the 2nd prize for Process in Science and Technology of Shanghai in 1999, China. He is currently an associate professor in School of Software, Shanghai Jiao Tong University. His research interests include Product Modeling and Automation Design.

Ailing Liu was born in 1976; she received her BS degree in School of Automatic Control department in 1998 from Huazhong University of Science and Technology, China. She received her MS degree in School of Software Engineering in 2010 from Shanghai Jiao Tong University, China. She is currently an Assistant Minister in IT department of Shanghai Waigaoqiao Shipbuilding Corporation.

Call for Papers and Special Issue Proposals

Aims and Scope.

Journal of Multimedia (JMM, ISSN 1796-2048) is a scholarly peer-reviewed international scientific journal published bimonthly, focusing on theories, methods, algorithms, and applications in multimedia. It provides a high profile, leading edge forum for academic researchers, industrial professionals, engineers, consultants, managers, educators and policy makers working in the field to contribute and disseminate innovative new work on multimedia.

The Journal of Multimedia covers the breadth of research in multimedia technology and applications. JMM invites original, previously unpublished, research, survey and tutorial papers, plus case studies and short research notes, on both applied and theoretical aspects of multimedia. These areas include, but are not limited to, the following topics:

- Multimedia Signal Processing
- Multimedia Content Understanding
- Multimedia Interface and Interaction
- Multimedia Databases and File Systems
- Multimedia Communication and Networking
- Multimedia Systems and Devices
- Multimedia Applications

JMM EDICS (Editors Information Classification Scheme) can be found at <http://www.academypublisher.com/jmm/jmmedics.html>.

Special Issue Guidelines

Special issues feature specifically aimed and targeted topics of interest contributed by authors responding to a particular Call for Papers or by invitation, edited by guest editor(s). We encourage you to submit proposals for creating special issues in areas that are of interest to the Journal. Preference will be given to proposals that cover some unique aspect of the technology and ones that include subjects that are timely and useful to the readers of the Journal. A Special Issue is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

The following information should be included as part of the proposal:

- Proposed title for the Special Issue
- Description of the topic area to be focused upon and justification
- Review process for the selection and rejection of papers.
- Name, contact, position, affiliation, and biography of the Guest Editor(s)
- List of potential reviewers
- Potential authors to the issue
- Tentative time-table for the call for papers and reviews

If a proposal is accepted, the guest editor will be responsible for:

- Preparing the "Call for Papers" to be included on the Journal's Web site.
- Distribution of the Call for Papers broadly to various mailing lists and sites.
- Getting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Instructions for Authors.
- Providing us the completed and approved final versions of the papers formatted in the Journal's style, together with all authors' contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

Special Issue for a Conference/Workshop

A special issue for a Conference/Workshop is usually released in association with the committee members of the Conference/Workshop like general chairs and/or program chairs who are appointed as the Guest Editors of the Special Issue. Special Issue for a Conference/Workshop is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

Guest Editors are involved in the following steps in guest-editing a Special Issue based on a Conference/Workshop:

- Selecting a Title for the Special Issue, e.g. "Special Issue: Selected Best Papers of XYZ Conference".
- Sending us a formal "Letter of Intent" for the Special Issue.
- Creating a "Call for Papers" for the Special Issue, posting it on the conference web site, and publicizing it to the conference attendees. Information about the Journal and Academy Publisher can be included in the Call for Papers.
- Establishing criteria for paper selection/rejections. The papers can be nominated based on multiple criteria, e.g. rank in review process plus the evaluation from the Session Chairs and the feedback from the Conference attendees.
- Selecting and inviting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Author Instructions. Usually, the Proceedings manuscripts should be expanded and enhanced.
- Providing us the completed and approved final versions of the papers formatted in the Journal's style, together with all authors' contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

More information is available on the web site at <http://www.academypublisher.com/jmm/>.

(Contents Continued from Back Cover)

A Watermarking Technique based on the Frequency Domain <i>Huang-Chi Chen, Yu-Wen Chang, and Rey-Chue Hwang</i>	82
Image Copy-Move Forgery Detecting Based on Local Invariant Feature <i>Li Jing and Chao Shao</i>	90
OWL-S based Service Composition of Threedimensional Geometry Modeling <i>Jiangning Yu, Hongming Cai, Fenglin Bu, and Ailing Liu</i>	98
