SUPERVISED SEMANTIC SCENE CLASSIFICATION BASED ON LOW-LEVEL CLUSTERING AND RELEVANCE FEEDBACK

A. DORADO, D. DJORDJEVIC, AND E. IZQUIERDO

Electronic Engineering Department
Queen Mary, University of London, Mile End Road, London
E1 4NS, UK
E-mail: ebroul.izquierdo@elec. qmul.ac.uk

W. PEDRYCZ

Department of Electrical and Computer Engineering
University of Alberta, ECERF, Edmonton, AB
T6G 2G7,Canada
E-mail: pedrycz@ee.ualberta.ca

A framework for semantic-based scene classification using relevance feedback is presented. The semantic component casts the classifier within a framework of the supervised –or *learning-from-examples*– paradigm. Selection of suitable examples and labeling training patterns imposes a certain burden on the user that increases with the complexity of the ontology involved in the scene interpretation. The proposed framework involves an on-line clustering whose intent is to create "natural" groups of patterns extracted from the scenes. The user adds some domain knowledge by labeling a number of randomly selected samples. Relevance feedback is incorporated to reinforce the training of the classifier in a 'learning with a critic' mode. To tackle the stability/plasticity dilemma that rises in changing the clusters arrangement, an intermediate structure is used to organize the patterns into semantically meaningful groups. The framework shows promising results and alleviates some of the drawbacks present when exploiting mechanisms of partial supervision when dealing with scene classification.

1. Introductory Remarks

The rapid growth in consumer-oriented electronic technologies, e.g. digital cameras, camcorders and mobile phones, along with the expansion in networking is facilitating production and consumption of impressive amounts of digital information. It also brings a change in the way people process such information. The challenge is in incorporating mechanisms to resemble the way humans make decisions based on how they interpret and what they perceive. This challenge has captured the attention of researchers in computer vision, pattern recognition, and other related fields in the last decades. The efforts are focused in adding knowledge to the multimedia content enabling more 'intelligent' processing.

Traditionally, proposed methods are used to designate a passage from visual primitives to human understanding of the multimedia content and to provide a way that a computer can execute the process [20]. This bottom-up approach relies completely on similarity at the lowest level. However, it is well known that two objects can be similar in their low-level primitives but semantically dissimilar to a human observer. This is a drawback in propagating interpretations using only low-level vision.

On the other hand, propagation based only on high-level similarity (top-down approach) puts a heavy burden on the user's shoulders and has undesirable effects in the overall system performance, which normally goes down and is stalled at certain point for lacking of information to make decisions without the user.

Hybrid approaches to go from the bottom to the top and in the opposite way are the foundation of the critical paradigm of "bridging the semantic gap." [6] Some of the current systems, e.g. content-based image/video retrieval systems (CBIR), use an *a-priori* training mode whence the system is trained in a supervised method. This method though effective presents some shortcomings in terms of efficiency. Here is when relevance feedback starts playing an important role allowing *a-posteriori* training modes and facilitating system's adaptation.

Although the semantic component places the systems into the supervised – or *learning-from-examples* paradigm, methods applied on low-level primitives allow a reduction of the user dependency. With this in mind and thinking on the feasibility of learning from human perception and understanding, a framework for semantic-based scene classification using relevance feedback is presented. The paper is organized as follows: Section 2 introduces the problem of narrowing the semantic gap in content-based scene classification; Sections 3 to 5 describe the proposed framework and the relevance feedback component; Section 6 gives a summary of selected experimental results; Concluding remarks and references are covered in Section 7 and 8, respectively.

2. The Scene Content Classification Problem

Visual primitives such as color and texture are extracted automatically from digital images using some measurements at pixel level and afterwards they become structured into feature vectors (patterns). Because they are generated from raw data at the lowest possible level of abstraction, these vectors are also known as low-level features. Formally, they are regarded as n-dimensional vectors in some n-dimensional space of reals, that is $\Re^n$).

Pattern recognition (PR) is based on the internal structures of the feature vectors [5]. In particular, PR techniques can help organize the patterns into "natural" clusters. The interpretation of these clusters may relate to some classification or description task pertaining to the relevant to the scene content, e.g. landscape, cityscape [8][22][21].

Cluster interpretations, or semantic categories, are associated with linguistic concepts at the highest level of abstraction. They are normally represented in the real domain ($\Re$) using either scalars or vectors. Categories are also referred to as high-level features. The category labels correspond to the concept tag. The category itself, attached to a content interpretation, denotes the meaning of the concept.

The recognizer performs a certain mapping from a starting structured data to a final structured data ($\Re^n \mapsto \Re$). However, the nature of the problem demands an extension to deal with the underneath subjectivity and fuzziness of the human interpretation [17]. That extension allows mapping from the final structured data to the starting one ($\Re \mapsto \Re^n$).

Consequently, the classification problem is stated as: "Find an intermediate structure to map from/to the starting structured data (natural clusters) to/from the final structured data.

3. A Framework for Semantic-Based Scene Classification

Making use of feature vectors provided by the feature extractor, a classifier is aimed at the assignment of objects to categories. to assign certain objects to a category [5]. Perceptually separable visual primitives are clustered either within some unsupervised (automatic) or with partially supervised (semi-automatic) pattern recognition mode. The classifier performs well (high classification accuracy) when cluster assignment is similar to the expected categorization.

An enhancement is introduced involving concept-wise human understanding. The interpretation of the human observer is captured labeling randomly images. Relevance feedback is used to present similar images to the labeled ones in order to receive inputs from the user regarding their relevance or irrelevance in terms of conceptual similarity. Figure 1 depicts an overall flowchart of the system.

The system displays a randomly set of images and the user collects a number of samples for each defined category; these samples are used as “seeds” to populate the conceptual clusters. Natural clusters are instantiated applying a clustering algorithm.

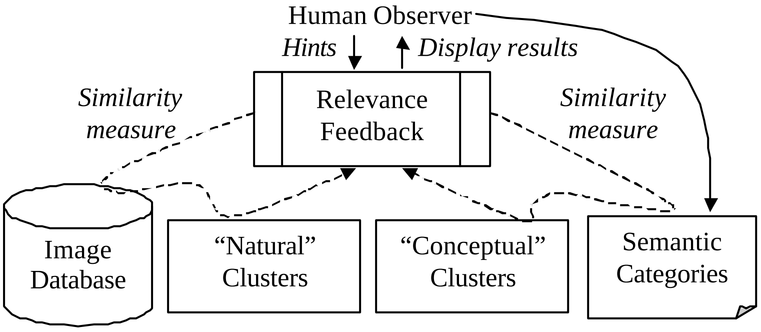


Figure 1. System flowchart.

The system uses the ascribed seeds to find the closer prototypes based on perceptual similarity. A search space is built with the natural clusters whence the similar prototypes belong to. Then, the classifier predicts positive examples for the category from the unlabelled images within the search space. Figure 2 illustrates the approach used to find positive samples for a category in order to train the classifier. The training stage is summarized below.

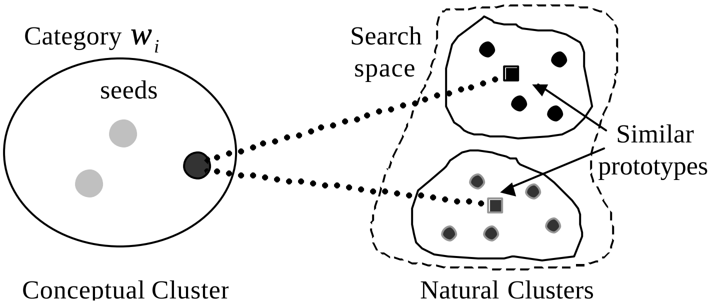


Figure 2. Finding positive samples for a category.

The user is asked to provide feedback indicating relevant images found among the positive examples. New positive examples are collected and used to update the seeds. These “seeds” can be also used to tune the “natural” clusters running partially supervised clustering [16][18][1]. The process of finding samples with the user-in-the-loop is repeated until relevant images are seldom found. Next, the learning continues with another category.

4. The Relevance Feedback Component

For the purpose of capturing both human perception of low-level features as well as human reasoning for conceptualizing at the highest level of abstraction relevance feedback approach is introduced.

Various relevance feedback algorithms have been proposed in the last decade as an integral part of content-based image retrieval systems. As mentioned in [10] early image relevance feedback methods were based on heuristic techniques with empiric parameter adaptation similarly as in weighting methods from text-based retrieval. The reasoning behind this idea is that the feature providing the most compact clustering of relevant images and separation of relevant from irrelevant images gets a higher weigh [19][3].

In modern content-based retrieval there is a significant effort to integrate support vector machine (SVM) into relevance feedback as a supervised learning method. SVMs describe hyperplanes in the feature space that separate classes and not the classes themselves [11][9]. SVMs belong to discriminative classification models that are not primarily concentrated on estimating the correct distribution of relevant and irrelevant data but rather on determining the boundaries between classes.

One of the directions modern relevance feedback techniques are heading considers solving a two-class (relevant/irrelevant) problem or even multi-class problem using a modification of discriminative analysis (DA) [23][24]. There are number techniques integrating neural network learning approaches and relevance feedback systems. For instance, in [12] the system uses hierarchical structure of Tree-structured Self-Organizing Map (TS-SOM) and in [25] a fuzzy relevance feedback approach based on radial basis function (RBF) neural networks is introduced.

The proposed framework uses SVMs and employs kernel-learning methods to optimize the non-linear mapping introduced with kernels for a better correspondence to the chosen feature representations [7][14][2].

5. SVM Classifier: Selecting the Kernel

Several images descriptors are combined in order to improve the effectiveness of the classifier. It rises the need of using appropriate distances for each descriptor as norms within the RBF kernel. Then, the kernel within SVMs has the following form:

$$K_{Gaussian}(\mathbf{x},\mathbf{y}) = \exp\left(-d(\mathbf{x},\mathbf{y}) \big/ 2s^{2}\right) \tag{1}$$

Where **x** and **y** denotes image descriptors. The distance $d(\mathbf{x}, \mathbf{y})$ is a linear combination of dynamically weighted (w_i) and normalized distances ($\bar{d}_i$) for each descriptor as follows:

$$d(\mathbf{x}, \mathbf{y}) = \sum_i w_i \bar{d}_i(\mathbf{x}, \mathbf{y}) \tag{2}$$

Weights calculation relies on the assumption that a particular descriptor resembling somehow the user preferences gets higher weigh. Thus the weights for each descriptor are determined as inverse of variance over all positive examples given by the user.

There is not assurance that the new kernel satisfies the Mercer's condition, guaranteeing kernels to be real inner products. Though it is possible to apply the SVM to kernels that do not satisfy such condition (cf. [4]). Besides, there is not longer guarantee that the optimal hyper plane maximizes the margin.

The combination of descriptors in a common feature space requires normalization to ensure equal emphasis for each descriptor. The overall mean value, m_i, and standard deviation, s_i, are calculated by representing the distribution of distances for all descriptors over the entire database. Mean and variance are then applied to normalize the appropriate individual distances within the linear combination used as a new norm for the generalized Gaussian kernel:

$$\bar{d}_i = \frac{d_i - m_i}{3s_i} \tag{3}$$

6. Experimental Studies

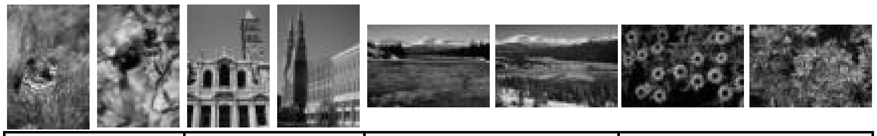
Experiments were conveyed with imagery selected from Corel stock gallery (corel.com). 1200 photographs were grouped into four categories: animal (dogs, subsea, and wl_rare); City view (ny_city, Ottawa, and rome); Landscape (autumn, can_west, and Yosemite); and Vegetation (perennial, plants, and usgarden). 720 (60%) of the images were used for training and 480 (40%) for testing.

The feature vectors are built concatenating three MPEG-7 descriptors namely color structure, edge histogram, and homogeneous texture. Each descriptor has a particular syntax and semantics (cf. [15]).

Table 1 presents accuracies achieved by the two-class and the multi-class classifier. The classifier trained with positives images from category animal presents higher misclassification of images belonging to class landscape. Confusion rates also report an increased misclassification between categories city view-landscape, landscape-vegetation, and vegetation-animal. Some samples of correctly classified and misclassified images are given in Figures 3 and 4.

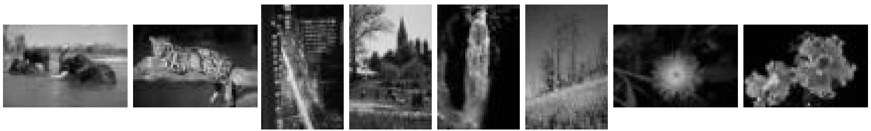
Table 1. Classifiers performance (%).

Animal	*City view*	*Landscape*	*Vegetation*	*Multi-class*
74.38	87.29	74.18	87.29	70.26



(a) Animal | (b) City view | (c) Landscape | (d) Vegetation

Figure 3. Samples of images correctly classified.



(a) Animal | (b) City view | (c) Landscape | (d) Vegetation

Figure 4. Samples of misclassified images.

Figure 5 shows some images that do not match any of the semantic categories and affected the overall performance of the multi-class classifier.

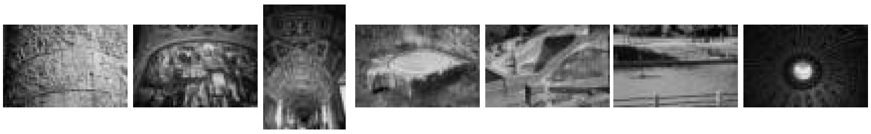


Figure 5. Samples of images that do not match any of the semantic categories.

7. Conclusions

A framework to classify scenes based on perception and interpretations of their content is presented. The approach seeks to narrow the gap rather than bridge it. It uses structures to shift low-level data towards high-level information. The learning stage applies relevance feedback to combine a-priori and a-posteriori training modes reducing the burden of the user and facilitating system adaptation.

Acknowledgments

Support from the Natural Sciences and Engineering Research Council (NSERC) and Canada Research Chair (W. Pedrycz) is gratefully acknowledged.

Part of the research leading to this paper was done within the framework of the project AceMedia – Integrating knowledge, semantics and content for user-centred intelligent media services (www.acemedia.org).

References

1. A. M. Bensaid et al. *Pattern Recognition*. 29, 859 (1996).
2. M. R. Boutell, J. Luo, X. Shen, and C. M. Brown. *Pattern Recognition*. **37**, 1757 (2004).
3. A. Celentano and S. Chiereghin. *Comp. Science Series*, **CS-99-10** (1999).
4. O. Chapelle et al. *IEEE Trans. on Neural Networks*. **10**, 1055 (1999).
5. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. 2nd Edition, Wiley-Interscience (2001).
6. C. Dorai and S. Venkatesh. *IEEE Multimedia*. **10**, 15 (2003).
7. S. R. Gunn. *Technical report*, University of Southampton. (1997).
8. M. M. Gorkani and R. W. Picard. *IEEE IAPR*. **1**, 459 (1994).
9. P. Hong, Q. Tian, and T. S. Huang. *IEEE ICME*. **2**, 1119 (2000).
10. T.S. Huang and X.S. Zhou. *IEEE ICIP*. **3**, 2 (2001).
11. F. Jing, M. Li, Hong-Jiang Zhang, and B. Zhang. *IEEE Trans. on Circuits and Systems for Video Technology*. **14**, 672 (2004).
12. M. Koskela, J. Laaksonen, and E. Oja. *CIVR*. **3115**, 508 (2004).
13. M. Koskela, J. Laaksonen, and E. Oja. *SCIA*. 579 (2001).
14. K.-R. Müller et al. *IEEE Neural Networks*. **12**,181 (2001).
15. B.S. Manjunath, J.-R. Ohm, V.V. Vasudevan, and A. Yamada. *IEEE Trans. on Circuits and Systems for Video Technology*. **11**, 703 (2001).
16. W. Pedrycz. *Pattern Recognition Letters*. **3**, 13 (1985).
17. W. Pedrycz. *Pattern Recognition*. **23**, 121 (1990).
18. W. Pedrycz and J. Waletzky. *IEEE Trans. on Systems, Man, and Cybernetics – Part B: Cybernetics*. **27**, 787 (1997).
19. Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. *IEEE Trans. on Circuits and Systems for Video Technology*. **8**, 644 (1998).
20. J. C. Simon. *Pattern Recognition*. **7**, 117 (1975).
21. A. Vailaya et al. *IEEE Trans. on Image Processing*, **10**, 117 (2001).
22. A. Vailaya, A. Jain, and H.-J. Zhang. *IEEE Workshop on Content-Based Access of Image and Video Libraries*. 3 (1998).
23. Q. Tian, Y. Wu, and T. S. Huang. *IEEE ICME*. **1**, 299 (2000).
24. Y. Wu, Q. Tian, and T. S. Huang. *IEEE ICPR*. **1**, 21 (2000).
25. K. Wu and K.-H. Yap. *IEEE ICICS-PRM*. **3**, 1595 (2003).