Evaluating Appearance Models for Recognition, Reacquisition, and Tracking

Doug Gray Shane Brennan Hai Tao

University of California, Santa Cruz
1156 High St., Santa Cruz, CA 95064
{*dgray, shanerb, tao*}*@soe.ucsc.edu*

Abstract

Traditionally, appearance models for recognition, reacquisition and tracking problems have been evaluated independently using metrics applied to a complete system. It is shown that appearance models for these three problems can be evaluated using a cumulative matching curve on a standardized dataset, and that this one curve can be converted to a synthetic disambiguation rate for single camera tracking or a synthetic reacquisition rate for cross camera tracking. A challenging new dataset for viewpoint invariant pedestrian recognition (VIPeR) is provided as an example. This dataset contains 632 pedestrian image pairs from arbitrary viewpoints. Several baseline methods are tested on this dataset and the results are presented as a benchmark for future appearance models and matching methods.

1. Introduction

Recognition, reacquisition and tracking are three of the most important topics in surveillance research today. The models used to represent objects in a surveillance system can usually be decomposed into three parts, an appearance model, a spatial model, and a temporal model. In a typical scenario, a tracking system will have access to all three parts, but will rely heavily on the spatial and temporal information. A reacquisition system will rely mostly on the appearance model, using the spatial and temporal information primarily to rule out unlikely matches. In contrast, a recognition system will only have access to the appearance model. Since the appearance information is largely independent of the spatial and temporal information, one can expect that an appearance model which performs well for the recognition problem will also perform well for the reacquisition and tracking problems. For this reason, we propose using recognition performance to characterize the performance of an appearance model for all three problems.

One of the most important tasks in surveillance research is the tracking of humans through a network of cameras. To evaluate an appearance model for this problem, a new dataset is provided which consists of pedestrian image pairs.

Figure 1: Some examples from our pedestrian dataset. Each column is one of 632 same-person example pairs. Note the wide range of viewpoint, pose, and illumination changes.

We define a pedestrian as a human with a restricted pose (upright). Each pair is an image of the same pedestrian taken from a different camera, under different viewpoint, pose and lighting conditions. We will refer to this dataset as the viewpoint invariant pedestrian recognition (VIPeR) dataset.[1] A few example image pairs are shown in Figure 1, note the differences in viewpoint, pose and lighting. In section 2 we provide a brief survey of some of the appearance models used in real systems, and evaluate many of them on our dataset.

Our proposed recognition methodology is motivated by the difficulty of performing recognition on a large pedestrian dataset without temporal information. Given a single image, the chance of choosing the correct match is inversely proportional to the size of the dataset. For this reason, we believe the cumulative matching characteristic (CMC) curve is the proper performance evaluation metric. This metric shows how performance improves as the number of requested images increases. It is shown that with some reasonable assumptions about the distribution of pedestrian appearances we can convert a CMC curve to a synthetic reacquisition/disambiguation rate for the reacquisition and

[1]The VIPeR dataset is available for download at: http://vision.soe.ucsc.edu/?q=node/178

tracking problems.

2. Previous Work

2.1. Appearance Models

The current state of the art appearance models include templates, subspace projections, manifolds, bags of features, constellations of features, exemplars, prototypes and histograms, among others. Template methods are the simplest but fail on pose or viewpoint changes. This problem can be alleviated with flexible matching and alignment [14], but this will only work for rigid objects with moderate viewpoint changes. Other more complicated template methods such as similarity templates [27] show some promise but have very high storage and computation requirements. Subspace methods may be sufficient to model pose and viewpoint changes [4], but the full subspace of complicated non-rigid objects like pedestrians is far too large to represent accurately. Manifold recognition techniques [22] have similar problems with dimensionality. Bag of feature methods use a collection of small features for recognition. These features could be simple Haar like features [20], complex image patches[1] [10] or SIFT features [21]. These methods are somewhat robust to viewpoint and pose changes, but are better suited for categorization problems where object shape is more important. Constellation methods use features with relative spatial information to improve performance [11] over simple bag of feature methods. Exemplar based methods [25] have shown promise for the related problem of cross viewpoint recognition, however this is not the same as viewpoint invariant and has only been tested on rigid objects such as vehicles. Prototype embedding is a recent advancement in this direction [15]. Histograms are one of the most robust to radical pose and viewpoint change because of the lack of spatial information. For this reason they are the basis of the mean shift algorithm [6] for tracking, the earth movers distance [24] for image retrieval, and the histogram of oriented gradients for pedestrian detection [7].

Two of the main drawbacks of the histogram approach are the lack of spatial information and the trade off between high quantization errors with coarse binning and sparseness associated with fine binning. Several attempts have been made to save spatial information in a histogram format. One of the most successful is the color correlogram proposed by Huang [17], which saves color correlation as a function of distance. Also noteworthy are multi-resolution histograms [16], image signatures [24], and spatiograms [3].

In practice, histogram representations are often optimized to suit a particular application. One example is a localized histogram, which we define as a collection of histograms taken over different regions of an image. Typically these regions are defined by a user to correspond to meaningful semantic regions such as head, shirt, and pants. For example, in [23], these regions are chosen as the top $\frac{1}{5}^{\text{th}}$, middle and bottom $\frac{2}{5}^{\text{ths}}$ of each image. We refer to this representation as a hand localized histogram. It is not difficult to see why a few histograms taken over high and low regions of an image would be a better representation for a pedestrian with different colored shirt and pants. For a fixed viewpoint these regions can be identified automatically using a conditional color model [27].

2.2. Matching

One of the practical problems with comparing appearance models across wide viewpoint changes is the illumination change between two different scene locations and/or camera models. Javed *et al.* has shown how these changes can be compensated for by learning brightness transfer functions [18]. An alternate method is simply to choose coarse histogram color bins which are somewhat invariant to illumination such as the primary and secondary colors [23]. It has been shown that the drift patterns of these colors can be learned and used to devise a specialized distance metric for comparing these kinds of histograms [28].

In [12], correspondences are established between images pairs, either using interest points or a decomposable triangular graph model. Their results look promising, but their dataset contains only 44 unique individuals seen from mostly frontal viewpoints.

2.3. Viewpoint Invariance

Viewpoint invariant recognition is a challenging problem on many levels. Current solutions to this problem can be roughly divided into three basic categories. The the holistic approach, an object is considered as a whole. This category includes histogram matching and all of its variants, as well subspace and manifold appearance models. Most histogram variants attempt to save some degree of spatial information which is usually the biggest drawback of the histogram. On the other hand, it is the lack of spatial information that makes histograms invariant to viewpoint and pose in the first place.

The second category is the piecemeal or part based approach. For rigid objects this might mean pose estimation and correction followed by a holistic method [14]. However for non-rigid objects, recognition may be done on parts alone to allow for occlusion or other missing features. This also allows for the relationship between parts to be used for recognition as well [11].

The third category is full 3D reconstruction. This method is likely to provide the best performance for most object classes, but is very costly and the novelty of viewpoint invariance is lost if one is able to do away with the viewpoint all together.

Of course *viewpoint invariance* can be interpreted to mean anything from invariant to "less than perfect alignment" to invariant to radical viewpoint and scale changes. We define it as invariant to any angle an object is likely to be seen from. Our viewpoint invariant dataset was created to evaluate digital video surveillance appearance models, so we define this to mean invariant to any rotation in the ground plane and reasonable rotation in the other dimensions.

3. VIPeR Dataset

While there exist a great deal of data for problems such as face detection, face recognition and pedestrian detection, as of the time of publication, there are no publicly available data for viewpoint invariant pedestrian recognition. In [27] and [12], authors present results on pedestrian datasets which contain primarily frontal pedestrian images. While this is reasonable in some scenarios such as with a confined indoor camera network, many surveillance scenarios require the ability to track pedestrians in large, open environments such as public plazas and airport terminals. In these environments a pedestrian may be seen from any angle, this is the the primary motivation for the viewpoint invariant approach.

3.1. Viewpoint Variation

This degree of variability makes data collection and selection more difficult. In order to claim that a recognition model is viewpoint invariant, we must use a dataset which contains all possible views of the object class to be learned. Since there are many degrees of freedom, this could potentially mean a huge dataset. If we quantize the range of viewpoints into 45 degree segments, we have 8 same viewpoint pairs and $\binom{8}{2} = 28$ *different* viewpoint pairs. Since we are interested in building a dataset to test performance across viewpoints, only these different viewpoint pairs are considered. We can utilize symmetry to reduce this number to 10, but collecting this many views of a single individual is still impractical. Our approach is to obtain a single viewpoint pair from many unique individuals from a pair of disparate cameras.

An ideal dataset would be constructed by uniformly sampling viewpoint pairs of single individuals from this set of 10 pairs, possibly with some bias to pairs which are seen more frequently. We attempt to achieve this ideal, but have some bias toward front-side viewpoint pairs. The distribution of viewpoints can be seen in figure 2. The distribution of viewpoint changes can be see in figure 3

3.2. Other Variations

The data were collected at many different locations over the course of several weeks. The cameras were placed in different locations in an academic setting and subjects were notified of the presence of cameras, but were not coached or instructed in any way. The illumination between cameras was allowed to vary freely. The quality of the images varies; the video was compressed before processing, as a result the images have spatially sub sampled chrominance channels, as well as some minor interlacing and compression artifacts.

Each image was cropped and scaled to be 128x48 pixels. While this scaling often caused some distortion, recent studies suggest that this does not have a significant effect on the human visual system [26] and we have not observed any significant effect on our results. This also removes biometric information such as height, but we consider this to be separate from the appearance model because it requires spatial information to compute.

Viewpoint angle (larger \ smaller)	0	45	90	135
45	16			
90	241	47		
135	43	72	4	
180	103	53	50	3

Figure 2: The distribution of viewpoint angles in the VIPeR dataset.

Viewpoint angle disparity	Examples
45	70
90	363
135	96
180	103

Figure 3: The distribution of viewpoint angle changes.

4. Performance Evaluation Methodology

The performance of a viewpoint invariant recognition system can be difficult to quantify because recognition performance is dependent upon the size of the dataset. To illustrate this, consider the performance of random guessing on a dataset of size 50 (2%) vs. a dataset of size 1000 (0.1%). In this example, an algorithm with a 10% recognition rate would be useless on the first dataset, but indispensable on the later. We say indispensable becasue for a human operator, the problem is relativly easy for small datasets but increasingly difficult as the number of possible matches grows. Analyis of how performance metrics vary with the dataset size can be found in [19] and [13].

But what exactly does *recognition rate* mean? There are two ways of looking at this problem, we can cast it as a same-different detection problem, or consider it as a ranking

problem.

4.1. Recognition as Detection

Several researchers [5] [25] have modeled the problem with the use of same/different probabilities. This allows for performance evaluation with an ROC curve. However this is a poor performance metric for recognition problems because as the size of the data set grows toward infinity, the prior probability of two objects being the same approaches zero. This is because the number of matching pairs in a dataset grows linearly (n) with the dataset size while the number of negative examples grows quadratically ($\frac{1}{2}n(n-1)$). This problem has lead some [20] [25] to use only some of the negative examples for training or testing. While this might make sense when training on a very large dataset, the ability to change the ratio of positive and negative examples allows one to manipulate the test error rate which is not a desirable quality in an evaluation metric.

4.2. Recognition as Ranking

A better evaluation methodology is to consider recognition as a ranking problem. In this framework a ranking is induced on the elements of the dataset and the probability that the correct match has a rank equal to or less than some value is plotted over the size of the test set. This performance metric is known as the cumulative matching characteristic (CMC) curve, which is analogous to the ROC curve for detection problems.

While this performance metric is designed to evaluate recognition problems, by making some simple assumptions about the distribution of appearances in a camera network we can convert a CMC curve into a synthetic disambiguation or reacquisition rate for multiple object or multiple camera tracking respectively. We say synthetic because we are converting a performance metric of size N to one of size M, where M is the number of elements in a synthetic example.

Assume that a set of M pedestrians that enter a camera network are i.i.d. samples from some large testing dataset of size N. If these M pedestrians cross from one camera to another at the same time we have a reacquisition problem where we must find the correct matching configuration. For the moment let us ignore the spatial and temporal information as well as the one to one matching constraint. If the CMC curve for the matching function is given, we can calculate the probability that any of the M best matches is correct as follows:

$$\mathrm{SDR}(M) = \mathrm{SRR}(M) = \mathrm{CMC}(N/M) \qquad (1)$$

Where $\mathrm{CMC}(k)$ is the rank k recognition rate. For example if $M = 2$, which is a common case, and $N = 100$ the disambiguation rate will be equal to the rank 50 recognition rate. This also applies to disambiguating pedestrians within a single camera. An example of this methodology applied to the VIPeR dataset can be found in figures 4 & 5.

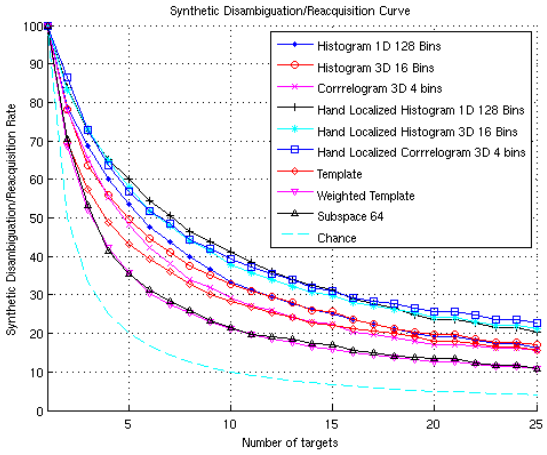

Figure 4: The synthetic disambiguation (SDR) or recognition (SRR) rate for $M = 1$ to 25. Details about each method can be found in section 5.

	M=2	M=3	M=4	M=5
Hist 1D	78.92	68.64	59.97	53.45
Hist 3D	78.23	63.67	55.98	49.81
Corel 3D	78.04	65.28	55.28	48.10
HL Hist 1D	83.70	72.66	65.19	60.03
HL Hist 3D	83.26	73.01	64.91	57.91
HL Corel 3D	86.46	72.72	63.73	56.71
Template	69.68	57.41	48.86	43.04
W Template	68.70	51.99	42.18	35.63
Subspace	70.41	53.16	41.30	35.44

Figure 5: The first few SDR/SRR values for each of the methods shown in figure 4.

Note that the SDR/SRR curve actually contains less information than the full CMC curve. However the information it contains is much more relevant to tracking and reacquisition problems where we rarely have more than a handful of possible matches.

In detection problems it has become common to use the area under the ROC curve as a single dimensional performance summary for a particular algorithm. We use the area under the CMC curve in an analogous manner to choose the best parameters for our baseline methods in section 5.

4.3. Cross Validation

Since recognition performance is dependent upon the size of the test set, it is important that all results reported use the same form of cross validation. Dietterich has shown that 5x2 cross validation represents a reasonable balance be-

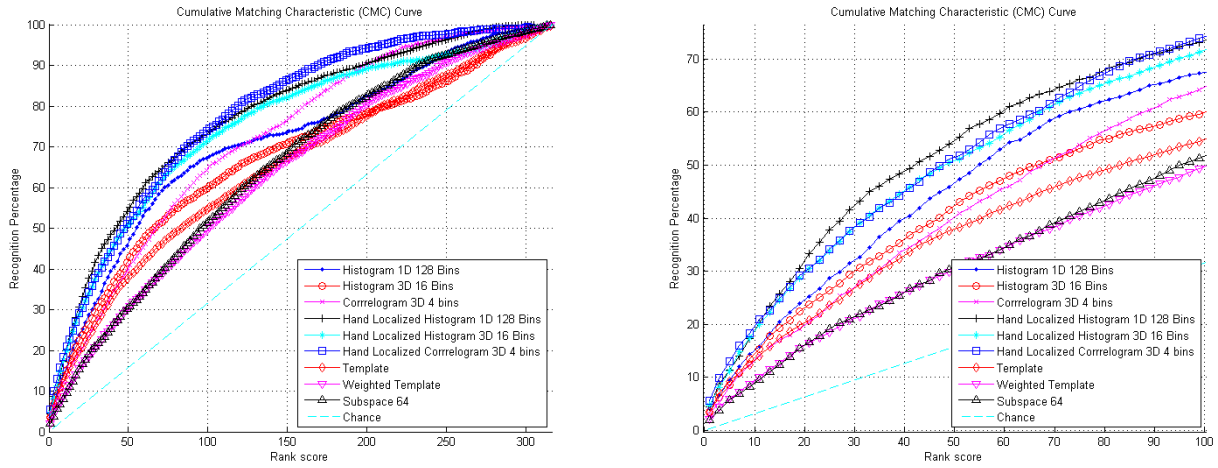

Figure 6: CMC curve for the best of each appearance model discussed (left), and a magnified version (right).

tween accuracy and experiment runtime [8]. When evaluating algorithm performance on the VIPeR dataset one should randomize both the training/test split and the order of the image pairs. Each image should also be mirrored at random to validate the assumptions under which the data was collected. At each round one must select one of the images from each pair as the *gallery* image and one as the *probe* image. In addition to averaging the results over the five rounds of two fold cross validation, one should also swap the gallery and probe images and average the result.

5. Results

5.1. Quantitative Results

To the best of our knowledge, there are no publicly available datasets for viewpoint invariant pedestrian recognition, thus there are no published quantitative results we can compare with. Several methods that have been used for related problems have been implemented to serve as a baseline. Since the set of all algorithms and configurations is much to large to display, only the best results of each class of algorithm and the parameters used are reported. The methods chosen for evaluation include templates, histograms, correlograms, subspace projections, and several variations of each. Several distance metrics were applied to each of the histogram based methods, but the Bhattacharyya distance provided the best results on average, and is used in all experiments here. The Bhattacharyya distance is a modified version of the Bhattacharyya coefficient [2] as discussed in [6]. Some details about each method are provided as follows:

- 1D Histograms were calculated over the entire image in the YCbCr colorspace. Channels were considered marginally and the number of bins was varied by powers of two. Optimal performance was achieved with 128 bins per channel.
- 3D Histograms were calculated in a similar manner as above except the channels were considered jointly. Optimal performance was achieved with 16 bins per channel.
- 3D Correlograms were calculated with 4 distance bins (1 3 5 9), and 2-5 color bins. Optimal performance was achieved with 4 color bins.
- Hand localized 1D/3D Histograms and Correlograms were calculated as above on the top fifth, middle and bottom two-fifths separately as was done in [23]. These three histograms were then concatenated before comparisons were made. The three regions were selected to correspond to head shirt and legs. The best parameters were the same as for the non-localized versions although their performance was much higher.
- Template methods use a simple sum of squared distance between two images. A weighted template applies a Gaussian weight to the center of each image to attempt to remove the effect of the background.
- Subspace methods were performed using principle component analysis on the training set. The data were then projected into the first n principal components and compared using the sum of squared distance. The parameter n was varied by powers of two. Performance for this method was the worst and did not increase significantly past $n = 64$ dimensions.

The CMC curves for the best performing of the above methods can be found in figure 6, while the SDR/SRR

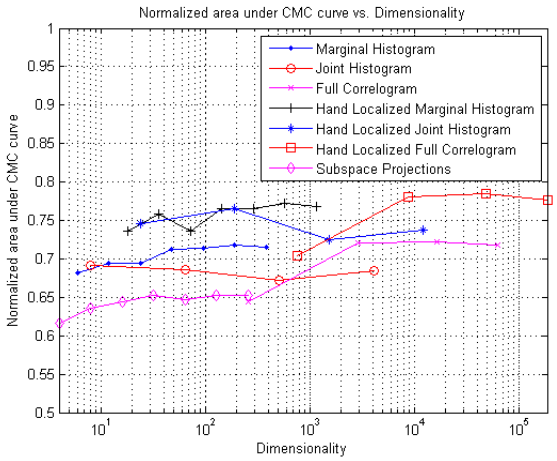

Figure 7: Plot of the area under the CMC curve as a function of the dimensionality of the representation.

curves can be found in figure 4. The first few values of the SDR/SRR curves can be found in figure 5.

In many practical applications, both the methods performance and efficiency must be considered. In figure 7 we compare the area under the CMC curve with the dimensionality of each representation. The dimensionality of the model is generally proportional to the runtime and storage requirements of the method. Thus we can use this plot to explore the efficacy of increasing the number of parameters in one model versus changing to another method. This illustrates the efficacy of using hand localized representations over increased binning for histograms. It also provides insight into how these simple methods can be improved. Since the hand localization was chosen empirically, we believe this could be optimized by applying some machine learning techniques to decompose the images into multiple regions.

6. Conclusions

This paper presents a new dataset for viewpoint invariant pedestrian recognition (VIPeR). A performance evaluation methodology is proposed for this new dataset and results for several baseline methods are provided and discussed. This dataset has been made public to promote future research on this challenging problem.

Our future work will involve the evaluation of several machine learning algorithms on the problem. Some possible directions include learning an energy based similarity metric such as in [5] or learning a compact representation with feature mining [9].

Regardless of how well an appearance model may perform on a static data set, in the real world pedestrians may alter their appearance by simply changing clothing. In this case recognition may become near impossible. However we are studying recognition primarily as a vehicle for evaluating the appearance models used in reacquisition and tracking. In these more constrained problems, a combination of spatial, temporal and appearance information should eventually be sufficient to solve any problem as well or better than a human operator.

References

[1] E. Bart, E. Byvatov, and S. Ullman. View-invariant recognition using corresponding object fragments. *Computer Vision. IEEE Computer Society Conference on*, pages 152–165, 2004.

[2] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc*, 35:99–109, 1943.

[3] S. Birchfield and S. Rangarajan. Spatiograms versus Histograms for Region-Based Tracking. *Computer Vision and Pattern Recognition. IEEE Computer Society Conference on*, 2, 2005.

[4] M. Black and A. Jepson. EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation. *International Journal of Computer Vision*, 26(1):63–84, 1998.

[5] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. *Computer Vision and Pattern Recognition. IEEE Computer Society Conference on*, 1, 2005.

[6] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. *Computer Vision and Pattern Recognition. IEEE Computer Society Conference on*, 2, 2000.

[7] N. Dalai and B. Triggs. Histograms of oriented gradients for human detection. *Computer Vision and Pattern Recognition. IEEE Computer Society Conference on*, 1, 2005.

[8] T. Dietterich. Approximate Statistical Test For Comparing Supervised Classification Learning Algorithms, 1998.

[9] P. Dollar, Z. Tu, H. Tao, and S. Belongie. Feature Mining for Image Classification. *Computer Vision and Pattern Recognition. IEEE Computer Society Conference on*, 2007.

[10] B. Epshtein and S. Ullman. Satellite Features for the Classification of Visually Similar Classes. *Computer Vision and Pattern Recognition IEEE Computer Society Conference on*, pages 2079–2086, 2006.

[11] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. *Computer Vision and Pattern Recognition. IEEE Computer Society Conference on*, 2, 2003.

[12] N. Gheissari, T. Sebastian, and R. Hartley. Person Reidentification Using Spatiotemporal Appearance. *Computer Vision and Pattern Recognition. IEEE Computer Society Conference on*, 2, 2006.

[13] P. Grother and P. Phillips. Models of large population recognition performance. *Computer Vision and Pattern Recognition. IEEE Computer Society Conference on*, 2.

[14] Y. Guo, S. Hsu, Y. Shan, and H. Sawhney. Vehicle fingerprinting for reacquisition & tracking in videos. *Computer Vision and Pattern Recognition. IEEE Computer Society Conference on*, 2, 2005.

[15] Y. Guo, Y. Shan, H. Sawhney, and R. Kumar. PEET: Prototype Embedding and Embedding Transition for Matching Vehicles over Disparate Viewpoints. *Computer Vision and Pattern Recognition. IEEE Computer Society Conference on*, 2007.

[16] E. Hadjidemetriou, M. Grossberg, and S. Nayar. Spatial information in multiresolution histograms. *Computer Vision and Pattern Recognition. IEEE Computer Society Conference on*, 1:702–709, 2001.

[17] J. Huang, S. Ravi Kumar, M. Mitra, W. Zhu, and R. Zabih. Spatial Color Indexing and Applications. *International Journal of Computer Vision*, 35(3):245–268, 1999.

[18] O. Javed, K. Shafique, and M. Shah. Appearance Modeling for Tracking in Multiple Non-overlapping Cameras. *Computer Vision and Pattern Recognition. IEEE Computer Society Conference on*, 2:26–33, 2005.

[19] A. Johnson, J. Sun, and A. Bobick. Using similarity scores from a small gallery to estimate recognition performance for larger galleries. *Analysis and Modeling of Faces and Gestures. IEEE International Workshop on*, pages 100–103, 2003.

[20] M. Jones and P. Viola. Face Recognition Using Boosted Local Features. *Mitsubishi Electric Research Laboratories Technical Report Number: TR2003-25. Date: April*, 2003.

[21] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[22] H. Murase and S. Nayar. Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision*, 14(1):5–24, 1995.

[23] U. Park, A. Jain, I. Kitahara, K. Kogure, and N. Hagita. ViSE: Visual Search Engine Using Multiple Networked Cameras. *Pattern Recognition. IEEE Computer Society Conference on*, pages 1204–1207, 2006.

[24] Y. Rubner, C. Tomasi, and L. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[25] Y. Shan, H. Sawhney, and R. Kumar. Vehicle Identification between Non-Overlapping Cameras without Direct Feature Matching. *Computer Vision. IEEE Computer Society International Conference on*, 1, 2005.

[26] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About. *Proceedings of the IEEE*, 94(11), 2006.

[27] C. Stauffer and E. Grimson. Similarity templates for detection and recognition. *Computer Vision and Pattern Recognition. IEEE Computer Society Conference on*, pages 221–228, 2001.

[28] G. Wu, A. Rahimi, K. Goh, C. Tsai, Y. Wu, E. Chang, and Y. Wang. Identifying color in motion in video sensors. *Computer Vision and Pattern Recognition. IEEE Computer Society Conference on*, pages 561–569, 2006.