APPEARANCE-BASED HUMAN GALLERY CONSTRUCTION FROM VIDEO

Kyongil Yoon
Computer Studies, College of Notre Dame of Maryland, Baltimore, MD, 21210, USA

Yaser Yacoob, David Harwood, Larry Davis
Computer Science Department, University of Maryland, College Park, MD, 20742, USA

Keywords: People recognition, Gallery construction, Appearance Modeling.

Abstract: An approach for constructing a dynamic gallery of people observed in a video stream is described. We consider two scenarios that require determining the number and identity of participants: outdoor surveillance and meeting rooms. In these applications face identification is typically not feasible due to the low resolution across the face. The proposed approach automatically computes an appearance model based on the clothing of people and employs this model in constructing and matching the gallery of participants. The appearance model uses *color/path-length* profile and a robust distance measure based on Kernel Density Estimation (KDE) and Kullback-Leibler (KL) distance, to evaluate similarity between people and add models to the gallery. A one-to-one constraint is enforced to correctly match instances to models at each frame. In the meeting room scenario we exploit the fact that the relative locations of subjects are likely to remain unchanged for the whole sequence.

1 INTRODUCTION

One aspect of video surveillance of indoor meetings involves matching a person against a gallery of known people. Such a gallery is tedious to construct manually; this paper describes an approach to automatically construct a gallery of participants based on clothing-appearance. The gallery directly supports the human identification task but it can also be used to answer questions such as how many people were observed, when each has appeared and how people interacted in video sequences.

We propose a method for building a gallery from a video clip based on clothing-appearance of people. We assume that people do not change clothing, although our method does tolerate localized appearance changes. We employ well-known approaches for human detection in video and focus on the modelling and matching of human appearance.

We consider two application areas: surveillance and meetings video. Here, it is difficult to employ faces for identification since the resolution across the face is too small and faces typically appear in off-frontal poses or profile views. Instead, we model the clothing of people and acquire quantitative models that support matching.

2 APPEARANCE MODELS

Over a short period of time, we assumed that the appearance of the person remains unchanged, except for small, local changes, for instance due to carried packages or illumination variation.

In (Nakajima et al., 2003), a full-body recognition system based on color and shape features has been suggested. They carried out recognition using support vector machine classifier on several features such as color histogram, normalized color histogram, combined histogram of shape and color, and local shape features. However, they did not combine spatial information with color as we do.

(Elgammal et al., 2002) segmented the human figure into three blobs and computed a separate color distribution for each blob. Specifically, the head, torso, and legs were segmented by assuming that the person appears in an upright pose. Although they separated the body into parts, most of the spatial informa-

Yoon K., Yacoob Y., Harwood D. and Davis L. (2007).
APPEARANCE-BASED HUMAN GALLERY CONSTRUCTION FROM VIDEO.
In *Proceedings of the Second International Conference on Signal Processing and Multimedia Applications*, pages 332-337
DOI: 10.5220/0002142103320337

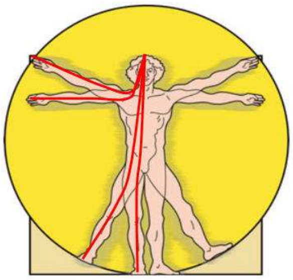

Figure 1: This is a simplified drawing of human body by Leonardo da Vinci. The red lines show shortest-path. The path-length is the distance from the top of the head to a given point on the path. The path-length to the end of hand or foot is relatively unchanged by the motion of the arms and legs.

tion is lost.

To overcome this problem, we introduced a simple, efficient feature, *path-length*, which represents the spatial information of a pixel with respect to a reference point on the person's body. The path-length is robust to changes in human posture and limb positions.

The *path-length* of a pixel is defined as the normalized length of the shortest path from the top center pixel (usually top of the head) inside a silhouette. Fig. 1 illustrates the idea of *path-length*.

In addition to *path-length*, clothing color information is employed to model appearance. The brightness (*Br*) defined as the sum of the three color components in (Alexander and Buxton, 2001), and two color proportions, *red* and *green* are used.

$$red = \frac{RED}{Br}, green = \frac{GREEN}{Br}.$$

3 MATCHING METRIC

The foreground region representing a person is used to construct an appearance model that is compared to models in the gallery. The distance between the current appearance and existing appearance models in the gallery determines if a new model should be added to the gallery or not.

3.1 Distance between Models

Our distance measure is computed based on kernel density estimation (KDE) and Kullback-Leibler (KL) distance. Kernel density estimation is a general non-parametric technique to estimate an underlying density using data points. In KDE, the probability for a given feature x is estimated as

$$\hat{f}(x) = \sum_i \alpha_i K(x - x_i),$$

where K is a kernel function centered at data points x_i, $i = 1...n$, and a_i are weighting coefficients. Typically, the Gaussian is used as a kernel function, and uniform weights are used, i.e., i = 1/n. Theoretically, suitable kernel density estimators converge to any density functions if enough samples are provided (Silverman, 1986) (Duda et al., 2000).

Assume that we are to compute the distance between Model $M = \{x_i | i = 1, \ldots, N_p\}$, where N_p is the number of data points in the appearance model, and current instance $I = \{y_i | i = 1, \ldots, N_q\}$, where N_q is the number of data points in the current instance. The estimated probability distribution of model M is

$$\hat{f}_M(x) = \sum_{i=1}^{N_p} \frac{1}{N_p} K_\sigma(x - x_i) \qquad (1)$$

and the distribution of the current instance, I is (2).

$$\hat{f}_I(x) = \sum_{i=1}^{N_q} \frac{1}{N_q} K_\sigma(x - y_i). \qquad (2)$$

The distance between the instance and the model can be thought of as the distance between two distributions represented by KDE, $\hat{f}_M(x)$ and $\hat{f}_I(x)$. The two most frequently used methods for comparing two distributions are Chi-Square test and Kolmogorov-Smirnov test (Press et al., 1988). Neither method is appropriate for our models. The Kolmogorov-Smirnov test is not suitable for our four dimensional model. The Chi-Square test involves dividing the data points into a number of bins; it is a good approximation when the number of bins is large ($\gg 1$), and number of events in each bin is large ($\gg 1$). However, for human appearance, the color distribution is very skewed, leading in many empty bins.

We instead use the Kullback-Leibler (KL) distance to compare $\hat{f}_M(x)$ and $\hat{f}_I(x)$. The KL distance is defined on two probability distributions in (Kapur and Kesavan, 1992), (Kullback and Leibler, 1951), (Cover and Thomas, 1991). For any given point, we can compute a pair of probabilities using the two estimated densities. Assume that there is a set of sample points, $S = \{s_i | i = 1, \ldots, n\}$, where n is the number of sample points. Then likelihood values for the sample points can be computed using

$$p_i = \hat{p}(s_i) = \frac{1}{N_p} \sum_{j=1}^{N_p} K_\sigma(s_i - x_j),$$

$$q_i = \hat{q}(s_i) = \frac{1}{N_q} \sum_{j=1}^{N_q} K_\sigma(s_i - y_j).$$

To compute the KL distance on those values, we normalize p_i and q_i as following

$$\hat{p}_i = \frac{p_i}{\sum_{j=1}^{n} p_j},\ \hat{q}_i = \frac{q_i}{\sum_{j=1}^{n} q_j}$$

The KL distance is defined as

$$d_{kl} = d(\hat{q}, \hat{p}) = \sum_{i=1}^{n} \hat{q}_i \log \frac{\hat{q}_i}{\hat{p}_i}$$

How to select S can be critical. To maximize the difference between p_i's and q_i's, it is best to use all the points in I; however, the computational cost can be prohibitive. Instead, by sampling points from I, we typically get equivalent results as long as the sampling process is reasonable. We sample points uniformly along *path-length* values. Practically, when we choose 100 points randomly spaced at 1% segments of path-length, the results are equivalent to using all the data points.

By examining the KL distance, we can measure how different two distributions are. However, because p_i and q_i are normalized, this can be problematical. For instance, when $\hat{f}_M(x)$ and $\hat{f}_I(x)$ are uniform distributions over different ranges, then all the p_i's are very low, and all the q_i's are very high. Although two distributions are quite different, after normalization $\hat{p}_i$ and $\hat{q}_i$ form almost identical distributions and d_{kl} is approximately and misleadingly 0.

To overcome this limitation, we introduce an additional distance measure which represents a quantitative difference between p_i's and q_i's as follows:

$$d_r = |1 - (\frac{\bar{p}}{\bar{q}})| \tag{3}$$

where $\bar{p} = (\sum p_i)/n$ and $\bar{q} = (\sum q_i)/n$.

3.2 Robust Distance Measure

Human appearance in video streams varies over time. In outdoor scenes, lighting, human pose variation and carried objects may lead to changes in the foreground region. To cope with such variations we employ a robust estimation norm that adjusts the weighting of points within the distance metric based on whether points are inliers or outliers.

For the robust estimation, we employ the general M-estimator of (Huber, 1977), which minimizes the objective function,

$$\sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho(y_i - \mathbf{x_i}^T \mathbf{b}) \tag{4}$$

where $\mathbf{x_i}$'s are independent variables, y_i's are data points, $\mathbf{b}$ is a coefficient vector, ρ is the influence function, and n is the number of data points.

If we define the weight function $\omega(e) = \rho'(e)/e$, and let $\omega_i = \omega(e_i)$. Then we need to solve the following equation to minimize (4)

$$\sum_{i=1}^{n} \omega_i (y_i - \mathbf{x_i^T b}) \mathbf{x_i^T} = 0 \tag{5}$$

In our approach, we define a new feature, δ_i using p_i and q_i for each sample point, s_i, :

$$\delta_i = \frac{|q_i - p_i|}{max(p_i, q_i)}$$

When the current instance is correctly matched to a model, most p_i's are similar to q_i's leading the δ_i's to be close to 0. On the other hand, when the instance and model are mismatched, most δ_i's will be greater than 0. The mean of δ_i will represent how well the current instance is matched to the model. We apply the robust fitting (5) to compute the robust mean of the δ_i's, μ; it can be written as

$$\sum_{i=1}^{n} \omega_i (\delta_i - \mu) = 0$$

Notice that weights are designed to minimize the influence of outliers. In other words, the weight of each data point depends on how far the point is from the mean. Data points near to the estimated mean get high weight. Points that are far from the mean have smaller weights.

We used the iteratively re-weighted least square (IRLS) method using the bisqaure weight function to solve the equation to get a robust mean as in (Coleman et al., 1980) and (Fox, 2002).

The final weights at the last iteration after the estimated mean converges were investigated to find inliers. Only data points with the weight greater than a certain threshold value are regarded as inliers. The two distances, d'_r and d'_{kl}, are recomputed using only inliers. Fig. 2 shows examples of outliers and inliers as determined using robust fitting method for a sample region that has been manually altered by changing its color.

4 SPATIAL ANALYSIS

Sometimes it is possible to improve the accuracy of the models in the gallery and the matching performance by utilizing the relative order of participants. We perform this as follows.

For each model, M_i, we compute an adjacency matrix, F_i that captures the frequency of spatial ordering among models. An adjacency matrix, F_i is $m \times n$, where n is the number of models and m indexes relative positions. For example, if N is the

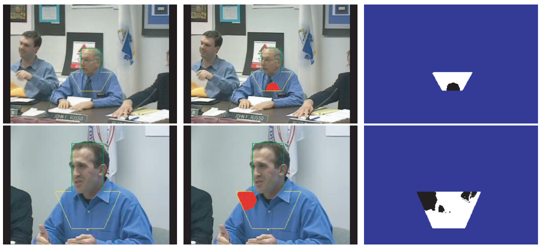

Figure 2: Detection of outliers. The image in the first column is the model image. Second column images are used as instances. To synthesize outliers, a 15% size block with red color pixels is created. In the third column the inliers and outliers are shown as white and black points, respectively.

maximum number of people in one frame and people are arranged in a ”linear” configuration, then $m = 2*(N-1)$.

To build the adjacency matrix F_i, all the frames which have a person matched to model M_i are employed. The (j,k)-th element of F_i is the frequency of model M_k at the relative horizontal position, $pos(j)$. $pos(.)$ is defined as

$$pos(j) = \begin{cases} j - \frac{m}{2} - 1 & \text{if } j < \frac{m}{2} \\ j - \frac{m}{2} & \text{otherwise} \end{cases} \tag{6}$$

The upper half of an adjacency matrix, F_i, represents the frequencies of models to the ”left” of M_i; the bottom the ”right” side.

The difference between adjacency matrices represents how similar two models are to each other. To compute the distance between adjacency matrices, the sum of absolute differences is used. Before computing d_{ij}, each F_i is normalized by the $max_{j,k}((F_i)_{j,k})$, so we have

$$d_{ij} = \sum_{k=1}^{n}\sum_{l=1}^{n} \left|(F_i)_{k,l} - (F_j)_{k,l}\right| \tag{7}$$

Fig. 3 shows the adjacency matrices from the experiment described in detail in section 5.2, 15 models were found after the first pass. Distances between adjacency matrices are computed, and pairs with distance less than a threshold can be merged into one.

5 EXPERIMENTS

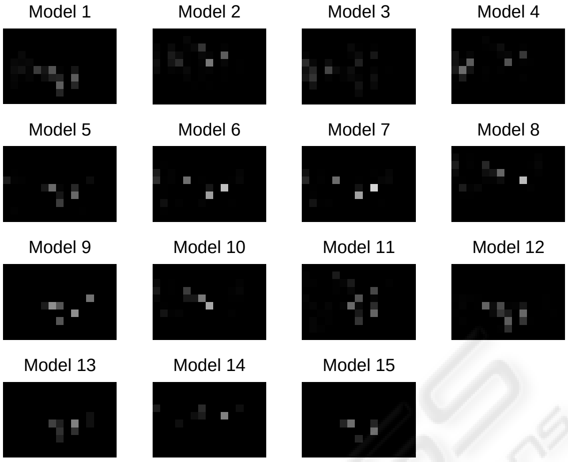

Figure 3: Adjacency matrices for 15 models in the experiment of section 5.2.

Figure 4: Sample frames of the full body gallery test.

We present two experiments. The first was conducted on four video clips collected at different locations and under different illumination conditions. The second experiment analyzes an 18 minute long video clip of a meeting. In this experiment, a face detection algorithm was used to determine an approximate torso area. In each experiment, we show the final gallery and the matching results based on the gallery.

The gallery construction process consists of two passes.

1. **Construct an initial gallery.** From an empty set, a gallery is built while processing all the frames. After this pass, the gallery has all the tentative models.
2. **Refine the gallery.** In this pass, redundant models are removed based on frequency and spatial analysis of the matching result, and a more compact and accurate gallery is built.

5.1 Full Body Gallery - Experiment 1

For this experiment, 1212 frames were collected from four different video clips. Three clips were outdoor video, and one clip was captured in a room monitoring people coming and going. The number of people in the test set is 12. We employed a background subtraction algorithm to detect the foreground regions. The detected regions are considered as full-body appearance of human. Fig. 4 shows some images in this test set.

After the first pass, we have 24 models in the gallery. The second pass uses the static gallery of the 24 models. In this experiment, most redundancy

Table 1: Matching result - Full body.

Gallery	Num of Models	Correct Match	Incorrect Match	Match Rate
Initial	24	1609	291	85.1%
Refined	16	1583	307	83.7%

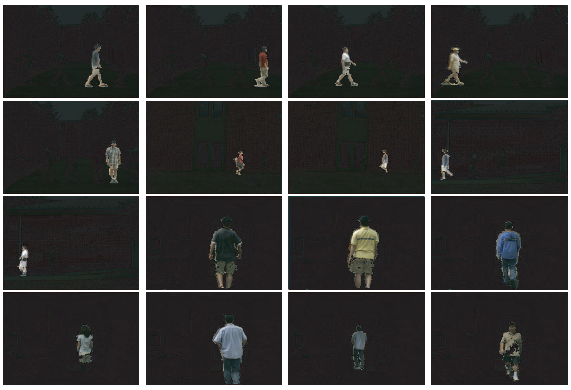

Figure 5: The final gallery built with a test set of Fig. 4. The number of models in the gallery is 16.

comes from the inaccuracies of human silhouettes created by background subtraction. After the second pass, we have a final gallery of 16 models as shown in Fig. 5. All 12 people have models. 2 people have two models ((M_3, M_9) and (M_{14}, M_{15})) and 1 person has 3 models respectively (M_1, M_7, M_8).

In this data set, 1890 foreground areas are detected from the 1212 frames. Using the the final gallery with 16 models, we could match 1583 regions correctly, while 307 are mismatched (83.7% success). When we use the 24 model gallery before removing redundant models, the number of correct matches is 1609 and 291 regions are not matched correctly (85.1% success). The representation power of the gallery is dependent on data set and foreground segmentation results. When using the same segmentation results, the final gallery has similar representation power compared to the gallery before redundant model removal (Table 1).

5.2 Upper Body Gallery - Experiment 2

An 18 minute long video clip which has 8 people is used for this experiment. Although the number of total frames is 32400, only one frame out of every five frames were processed. This video clip was captured in a meeting room, and people remain seated without position changes. The cameras pan and tilt as the meeting progresses, so that at any one time we see a different subset of the participants. Only the upper bodies of people are seen.

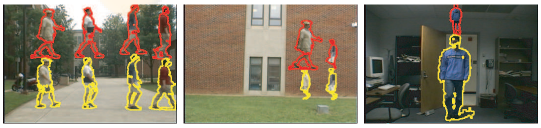

Figure 6: Some frames showing matching results with the final gallery.

Figure 7: Sample frames from the video clip used in upper body gallery test.

Table 2: Frequency of each model.

Model	M_1	M_2	M_3	M_4	M_5
Freq.	910	703	370	277	1945
Model	M_6	M_7	M_8	M_9	M_{10}
Freq.	426	1892	2997	9	3359
Model	M_{11}	M_{12}	M_{13}	M_{14}	M_{15}
Freq.	97	221	16	18	7

We employ a face detection algorithm to locate people (Viola and Jones, 2001). Based on the detected faces, the torso areas were computed and appearance matching was conducted. Since the relative positions between people remain unchanged for the entire clip, we perform the spatial analysis described in section 4. Several frames are shown in Fig. 7. The first pass constructed a 15 model gallery excluding false alarms from the face detector.

In the second pass, the spatial analysis of relative horizontal positions was carried out. The adjacency matrices of the 15 models were shown in Fig. 3.

Before calculating the differences between adjacency matrices, the total frequency for each model is used to eliminate some models. The total number of face occurrences is 13709, and some models have very low frequency. Table 2 shows the frequency of each model.

As seen in Table 2, M_9, M_{13}, M_{14}, M_{15} can be eliminated since their frequencies are very low. Next, by thresholding the differences between adjacency matrices, we select pairs of models, which can be merged into one.

The final gallery has 8 models. In the video clip, although there are nine people appearing, the ninth person shows only side view and she was not detected by the face recognition algorithm. The eighth person was not included in the gallery, and two models were found for the first person. Table 3 shows the gallery we constructed. The merged models are shown in parentheses. Fig. 9 shows some of the matching re-

Table 3: Final Gallery.

Person	Model
P_1	M_3,M_4
P_2	M_2
P_3	(M_1,M_{12})
P_4	(M_5,M_9,M_{11},M_{13},M_{15})
P_5	(M_6,M_7)
P_6	M_{10}
P_7	(M_8,M_{14})
P_8	NONE

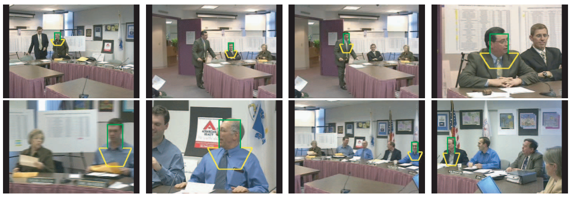

Figure 8: 8 models in the final gallery after the spatial analysis.

sults using the final gallery. To investigate the identification accuracy of matching, we randomly chose 100 frames which were found to have 210 face areas. Table 4 summarized the result. Just like in the experiment in Sec. 5.1, even with the smaller number of models the gallery shows the similar performance.

6 CONCLUSION AND FUTURE WORK

We proposed an approach for constructing a dynamic gallery of people from a video clip or a set of frame images based on appearance model using color/path-length profile. Kullback-Leibler distance is used to robustly compare models and a one-to-one constraint is enforced when more than one instance is present and matched in a frame. When the order of people rarely changes, the relative spatial order is analyzed and used to reduce the redundant models from the gallery.

There is trade-off between representation power and compactness of gallery. Using multiple key-frames to build a model can help to give more representation power to the models. One of our future work is to find an effective method to accumulate the information of multiple frames into one model.

Table 4: Matching result - Upper body.

Gallery	Num of Models	Correct Match	Incorrect Match	Match Rate
Initial	15	198	12	94.3%
Refined	8	194	16	92.4%

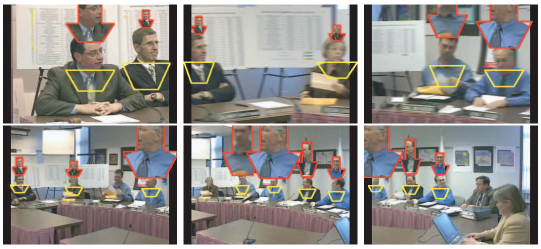

Figure 9: Some frames showing matching results with the final gallery in the second experiment.

REFERENCES

Alexander, D. C. and Buxton, B. F. (2001). Statistical modeling of colour data. *International Journal of Computer Vision*, 44(2):87–109.

Coleman, D., Holland, P., Kaden, N., Klema, V., and Peters, S. C. (1980). A system of subroutines for iteratively reweighted least squares computations. *ACM Trans. Math. Softw.*, 6(3):327–336.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. New York: Wiley.

Duda, R. O., Stork, D., and Hart, P. E. (2000). *Pattern Classification*. John Wiley and Sons Inc.

Elgammal, A., Duraiswami, R., Harwood, D., and Davis, L. S. (2002). Background and foreground modeling using non-parametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7):1151–1163.

Fox, J. (2002). Robust regression: Appendix to an r and s-plus companion to applied regression.

Huber, P. J. (1977). *Robust Statistical Procedures*. Society for Industrial and Applied Mathematics.

Kapur, J. N. and Kesavan, H. K. (1992). *Entropy Optimization Principles with Applications*. Academic Press.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

Nakajima, C., Pontil, M., Heisele, B., and Poggio, T. (2003). Full-body person recognition system. *Pattern Recognition*, 36(9):1997–2006.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1988). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall, New York.

Viola, P. A. and Jones, M. J. (2001). Rapid object detection using a boosted cascade of simple features. In *CVPR (1)*, pages 511–518.