Simultaneous Appearance Modeling and Segmentation for Matching People under Occlusion

Zhe Lin, Larry S. Davis, David Doermann, and Daniel DeMenthon

Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742, USA
{zhelin,lsd,doermann,daniel}@umiacs.umd.edu

Abstract. We describe an approach to segmenting foreground regions corresponding to a group of people into individual humans. Given background subtraction and ground plane homography, hierarchical part-template matching is employed to determine a reliable set of human detection hypotheses, and progressive greedy optimization is performed to estimate the best configuration of humans under a Bayesian MAP framework. Then, appearance models and segmentations are simultaneously estimated in an iterative sampling-expectation paradigm. Each human appearance is represented by a nonparametric kernel density estimator in a joint spatial-color space and a recursive probability update scheme is employed for soft segmentation at each iteration. Additionally, an automatic occlusion reasoning method is used to determine the layered occlusion status between humans. The approach is evaluated on a number of images and videos, and also applied to human appearance matching using a symmetric distance measure derived from the Kullback-Leiber divergence.

1 Introduction

In video surveillance, people often appear in small groups, which yields occlusion of appearances due to the projection of the 3D world to 2D image space. In order to track people or to recognize them based on their appearances, it would be useful to be able to segment the groups into individuals and build their appearance models. The problem is to segment foreground regions from background subtraction into individual humans.

Previous work on segmentation of groups can be classified into two categories: detection-based approaches and appearance-based approaches. Detection-based approaches model humans with 2D or 3D parametric shape models (e.g. rectangles, ellipses) and segment foreground regions into humans by fitting these models. For example, Zhao and Nevatia [1] introduce an MCMC-based optimization approach to human segmentation from foreground blobs. Following this work, Smith et al. [2] propose a similar trans-dimensional MCMC model to track multiple humans using particle filters. Later, an EM-based approach is

proposed by Rittscher et al. [3] for foreground blob segmentation. On the other hand, appearance-based approaches segment foreground regions by representing human appearances with probabilistic densities and classifying foreground pixels into individuals based on these densities. For example, Elgammal and Davis [4] introduce a probabilistic framework for human segmentation assuming a single video camera. In this approach, appearance models must first be acquired and used later in segmenting occluded humans. Mittal and Davis [5] deal with the occlusion problem by a multi-view approach using region-based stereo analysis and Bayesian pixel classification. But this approach needs strong calibration of the cameras for its stereo reconstruction. Other multi-view-based approaches [6][7][8] combine evidence from different views by exploiting ground plane homography information to handle more severe occlusions.

Our goal is to develop an approach to segment and build appearance models from a single view even if people are occluded in every frame. In this context, appearance modeling and segmentation are closely related modules. Better appearance modeling can yield better pixel-wise segmentation while better segmentation can be used to generate better appearance models. This can be seen as a chicken-and-egg problem, so we solve it by the EM algorithm. Traditional EM-based segmentation approaches are sensitive to initialization and require appropriate selection of the number of mixture components. It is well known that finding a good initialization and choosing a generally reasonable number of mixtures for the traditional EM algorithm remain difficult problems. In [15], a sample consensus-based method is proposed for segmenting and tracking small groups of people using both color and spatial information. In [13], the KDE-EM approach is introduced by applying the nonparametric kernel density estimation method in EM-based color clustering. Later in [14], KDE-EM is applied to single human appearance modeling and segmentation from a video sequence.

We modify KDE-EM and apply it to our problem of foreground human segmentation. First, we represent kernel densities of humans in a joint spatial-color space instead of density estimation in a pure color space. This can yield more discriminative appearance models by enforcing spatial constraints on color models. Second, we update assignment probabilities *recursively* instead of using a direct update scheme in KDE-EM; this modification of feature space and update equations results in faster convergence and better segmentation accuracy. Finally, we propose a general framework for building appearance models from occluded humans and matching them using full or partial observations.

2 Human Detection

In this section, we briefly introduce our human detection approach (details can be found in [16]). The detection problem is formulated as a Bayesian MAP optimization [1]: $\mathbf{c}^* = \arg\max_{\mathbf{c}} P(\mathbf{c}|I)$, where I denotes the original image, $\mathbf{c} = \{\mathbf{h}_1, \mathbf{h}_2, ...\mathbf{h}_n\}$ denotes a human configuration (a set of human hypotheses). $\{\mathbf{h}_i = (\mathbf{x}_i, \theta_i)\}$ denotes an individual hypothesis which consists of foot position $\mathbf{x}_i$ and corresponding model parameters θ_i (which are defined as the indices of part-

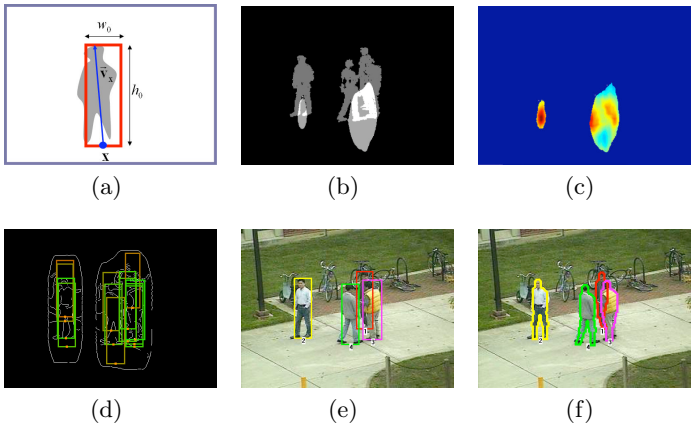

Fig. 1. An example of the human detection process. (a) Adaptive rectangular window, (b) Foot candidate regions $\mathbf{R}_{foot}$ (lighter regions), (c) Object-level likelihood map by hierarchical part-template matching, (d) The initial set of human hypotheses overlaid on the Canny edge map, (e) Human detection result, (f) Shape segmentation result.

templates). Using Bayes Rule, the posterior probability can be decomposed into a joint likelihood and a prior as: $P(\mathbf{c}|I) = \frac{P(I|\mathbf{c})P(\mathbf{c})}{P(I)} \propto P(I|\mathbf{c})P(\mathbf{c})$. We assume a uniform prior, hence the MAP problem reduces to maximizing the joint likelihood. The joint likelihood $P(I|\mathbf{c})$ is modeled as a multi-hypothesis, multi-blob observation likelihood. The multi-blob observation likelihood has been previously explored in [9][10].

Hierarchical part-template matching is used to determine an initial set of human hypotheses. Given the (off-line estimated) foot-to-head plane homography [3], we search for human foot candidate pixels by matching a part-template tree to edges and binary foreground regions hierarchically and generate the object-level likelihood map. Local maxima are chosen adaptively from the likelihood map to determine the initial set of human hypotheses. For efficient implementation, we perform matching only for pixels in foot candidate regions $\mathbf{R}_{foot}$. $\mathbf{R}_{foot}$ is defined as: $\mathbf{R}_{foot} = \{\mathbf{x}|\gamma_{\mathbf{x}} \geq \xi\}$, where $\gamma_{\mathbf{x}}$ denotes the proportion of foreground pixels in an adaptive rectangular window $W(\mathbf{x}, (w_0, h_0))$ determined by the human vertical axis $\overrightarrow{\mathbf{v}}_{\mathbf{x}}$ (estimated from the homography mapping). The window coverage is efficiently calculated using integral images. Then, a fast and efficient greedy algorithm is employed for optimization. The algorithm works in a progressive way as follows: starting with an empty configuration, we iteratively add a new, locally best hypothesis from the remaining set of possible hypotheses until the termination condition is satisfied. The iteration is terminated when the joint likelihood stops increasing or no more hypothesis can be added. Fig. 1 shows an example of the human detection process.

3 Human Segmentation

3.1 Modified KDE-EM Approach

KDE-EM [13] was originally developed for figure-ground segmentation. It uses nonparametric kernel density estimation [11] for representing feature distributions of foreground and background. Given a set of sample pixels $\{\mathbf{x}_i, i = 1, 2...N\}$ (with a distribution $\mathcal{P}$), each represented by a d-dimensional feature vector as $\mathbf{x}_i = (x_{i1}, x_{i2}..., x_{id})^t$, we can estimate the probability $\hat{P}(\mathbf{y})$ of a new pixel $\mathbf{y}$ with feature vector $\mathbf{y} = (y_1, y_2, ..., y_d)^t$ belonging to the same distribution $\mathcal{P}$ as:

$$\hat{p}(\mathbf{y} \in \mathcal{P}) = \frac{1}{N\sigma_1...\sigma_d} \sum_{i=1}^{N} \prod_{j=1}^{d} k(\frac{y_j - x_{ij}}{\sigma_j}), \tag{1}$$

where the same kernel function $k(\cdot)$ is used in each dimension (or channel) with different bandwidth σ_j. It is well known that a kernel density estimator can converge to any complex-shaped density with sufficient samples. Also due to its nonparametric property, it is a natural choice for representing the complex color distributions that arise in real images.

We extend the color feature space in KDE-EM to incorporate spatial information. This joint spatial-color feature space has been previously explored for feature space clustering approaches such as [12], [15]. The joint space imposes spatial constraints on pixel colors hence the resulting density representation is more discriminative and can tolerate small local deformations. Each pixel is represented by a feature vector $\mathbf{x} = (X^t, C^t)^t$ in a 5D space, $\mathbb{R}^5$, with 2D spatial coordinates $X = (x_1, x_2)^t$ and 3D normalized rgs color[1] coordinates $C = (r, g, s)^t$. In Equation 1, we assume independence between channels and use a Gaussian kernel for each channel. The kernel bandwidths are estimated as in [11].

In KDE-EM, the foreground and background assignment probabilities $\hat{f}^t(\mathbf{y})$ and $\hat{g}^t(\mathbf{y})$ are updated directly by weighted kernel densities. We modify this by updating the assignment probabilities *recursively* on the previous assignment probabilities with weighted kernel densities (see Equation 2). This modification results in faster convergence and better segmentation accuracy, which is quantitatively verified in [17] in terms of pixel-wise segmentation accuracy and number of iterations needed for foreground/background segmentation.

3.2 Foreground Segmentation Approach

Given a foreground regions $\mathbf{R}_f$ from background subtraction and a set of initial human detection hypotheses $(\mathbf{h}_k, k = 1, 2, ...K)$, the problem of segmentation is equivalent to the K-class pixel labeling problem. The label set is denoted as $\mathcal{F}_1, \mathcal{F}_2, ...\mathcal{F}_K$. Given a pixel $\mathbf{y}$, we represent the probability of pixel $\mathbf{y}$ belonging to human-k as $\hat{f}_k^t(\mathbf{y})$, where $t = 0, 1, 2...$ is the iteration index. The assignment probabilities $\hat{f}_k^t(\mathbf{y})$ are constrained to satisfy the condition: $\sum_{k=1}^{K} \hat{f}_k^t(\mathbf{y}) = 1$.

[1] $r = R/(R + G + B)$, $g = G/(R + G + B)$, $s = (R + G + B)/3$

Algorithm 1 Initialization by Layered Occlusion Model

initialize $R_0^0(\mathbf{y}) = 1$ for all $\mathbf{y} \in \mathbf{R}_f$
for $k = 1, 2, ...K-1$
- for all $\mathbf{y} \in \mathbf{R}_f$
- $\hat{f}_k^0(\mathbf{y}) = R_{k-1}^0(\mathbf{y}) e^{-1/2(Y-Y_{0,k})^t V^{-1}(Y-Y_{0,k})}$ and $R_k^0(\mathbf{y}) = 1 - \sum_{i=1}^{k} \hat{f}_i^0(\mathbf{y})$
endfor
set $\hat{f}_K^0(\mathbf{y}) = R_{K-1}^0(\mathbf{y})$ for all $\mathbf{y} \in \mathbf{R}_f$ and **return** $\hat{f}_1^0, \hat{f}_2^0, ..., \hat{f}_K^0$
where Y denotes the spatial coordinates of $\mathbf{y}$, $Y_{0,k}$ denotes the center coordinates of object k, and V denotes the covariance matrix of the 2D spatial Gaussian distribution.

Layered Occlusion Model. We introduce a layered occlusion model into the initialization step. Given a hypothesis of an occlusion ordering of detections, we build a layered occlusion representation iteratively by calculating the foreground probability map $\hat{f}_k^0$ for the current layer and its residual probability map R_k^0 for pixel $\mathbf{y}$). Suppose the occlusion order (from front to back) is given by $\mathcal{F}_1, \mathcal{F}_2, ...\mathcal{F}_K$; then the initial probability map is calculated recursively from the front layer to the back layer by assigning 2D anisotropic Gaussian distributions based on the location and scales of each detection hypothesis.

Occlusion Reasoning. The initial occlusion ordering is determined by sorting the detection hypotheses by their vertical coordinates and the layered occlusion model is used to estimate initial assignment probabilities. The occlusion status is updated at each iteration (after the $E-step$) by comparing the evidence of occupancy in the overlap area between different human hypotheses. For two human hypotheses $\mathbf{h}_i$ and $\mathbf{h}_j$, if they have overlap area $O_{\mathbf{h}_i,\mathbf{h}_j}$, we re-estimate the occlusion ordering between the two as: $\mathbf{h}_i$ occlude $\mathbf{h}_j$ if $\sum_{\mathbf{x}\in O_{\mathbf{h}_i,\mathbf{h}_j}} \hat{f}_i^t(\mathbf{x}) > \sum_{\mathbf{x}\in O_{\mathbf{h}_i,\mathbf{h}_j}} \hat{f}_j^t(\mathbf{x})$ (i.e. $\mathbf{h}_i$ better accounts for the pixels in the overlap area than $\mathbf{h}_j$), $\mathbf{h}_j$ occlude $\mathbf{h}_i$ otherwise, where $\hat{f}_i^t$ and $\hat{f}_j^t$ are the foreground assignment probabilities of $\mathbf{h}_i$ and $\mathbf{h}_j$. At each iteration, every pair of hypotheses is compared in this way if they have a non-empty overlap area. The whole occlusion ordering is updated by exchanges if and only if an estimated pairwise occlusion ordering differs from the previous ordering.

4 Partial Human Appearance Matching

Appearance models represented by kernel probability densities can be compared by information theoretic measures such as the Battacharyya Distance or the Kullback Leiber Distance for tracking and matching objects in video. Recently, Yu et al. [18] introduce an approach to construct appearance models from a video sequence by a key frame method and show robust matching results using a path-length feature and the Kullback-Leiber distance measure. But this approach only handles un-occluded cases.

Algorithm 2 Simultaneous Appearance Modeling and Segmentation for Occlusion Handling

Given a set of sample pixels $\{\mathbf{x}_i, i = 1, 2...N\}$ from the foreground regions $\mathbf{R}_f$, we iteratively estimate the assignment probabilities $\hat{f}_k^t(\mathbf{y})$ of a foreground pixel $\mathbf{y} \in \mathbf{R}_f$ belonging to $\mathcal{F}_k$ as follows:

Initialization : Initial probabilities are assigned by the layered occlusion model.

M − Step : (Random Pixel Sampling) We randomly sample a set of pixels (we use $\eta = 5\%$ of the pixels) from the foreground regions $\mathbf{R}_f$ for estimating each foreground appearances represented by weighted kernel densities.

E − Step : (Soft Probability Update) For each $k \in \{1, 2, ...K\}$, the assignment probabilities $F_k^t(\mathbf{y})$ are recursively updated as follows:

$$\hat{f}_k^t(\mathbf{y}) = c\hat{f}_k^{t-1}(\mathbf{y}) \sum_{i=1}^{N} \hat{f}_k^{t-1}(\mathbf{x}_i) \prod_{j=1}^{d} k(\frac{y_j - x_{ij}}{\sigma_j}), \tag{2}$$

where N is the number of samples and c is a normalizing constant such that $\sum_{k=1}^{K} \hat{f}_k^t(\mathbf{y}) = 1$.

Segmentation : The iteration is terminated when the average segmentation difference of two consecutive iterations is below a threshold:

$$\frac{\sum_k \sum_{\mathbf{y}} |\hat{f}_k^t(\mathbf{y}) - \hat{f}_k^{t-1}(\mathbf{y})|}{nK} < \epsilon, \tag{3}$$

where n is the number of pixels in the foreground regions. Let $\hat{f}_k(\mathbf{y})$ denote the final converged assignment probabilities. Then the final segmentation is determined as: pixel $\mathbf{y}$ belong to human-k, i.e. $\mathbf{y} \in \mathcal{F}_k, k = 1, ...K$, if $k = \arg\max_{k \in \{1,...K\}} \hat{f}_k(\mathbf{y})$.

Suppose two appearance models, a and b are represented as kernel densities in a joint spatial-color space. Assuming a as the reference model and b as the test model, the similarity of the two appearances can be measured by the Kullback-Leiber distance as follows [12][18]:

$$D_{KL}(\hat{p}^b || \hat{p}^a) = \int \hat{p}^a(\mathbf{y}) log \frac{\hat{p}^a(\mathbf{y})}{\hat{p}^b(\mathbf{y})} d\mathbf{y}, \tag{4}$$

where $\mathbf{y}$ denotes a feature vector, and $\hat{p}^a$ and $\hat{p}^b$ denote kernel pdf functions. For simplification, the distance is calculated from samples instead of the whole feature set. We need to compare two kernel pdfs using the same set of samples in the feature space. Given N_a samples $\mathbf{x}_i, i = 1, 2..., N_a$ from the appearance model a and N_b samples $\mathbf{y}_k, k = 1, 2..., N_b$ from the appearance model b, the above equation can be approximated by the following form [18] given sufficient samples from the two appearances:

$$D_{KL}(\hat{p}^b || \hat{p}^a) = \frac{1}{N_b} \sum_{k=1}^{N_b} log \frac{\hat{p}^b(\mathbf{y}_k)}{\hat{p}^a(\mathbf{y}_k)}, \tag{5}$$

$$\hat{p}^a(\mathbf{y}_k) = \frac{1}{N_a} \sum_{i=1}^{N_a} \prod_{j=1}^{d} k(\frac{y_{kj} - x_{ij}}{\sigma_j}), \ \hat{p}^b(\mathbf{y}_k) = \frac{1}{N_b} \sum_{i=1}^{N_b} \prod_{j=1}^{d} k(\frac{y_{kj} - y_{ij}}{\sigma_j}). \tag{6}$$

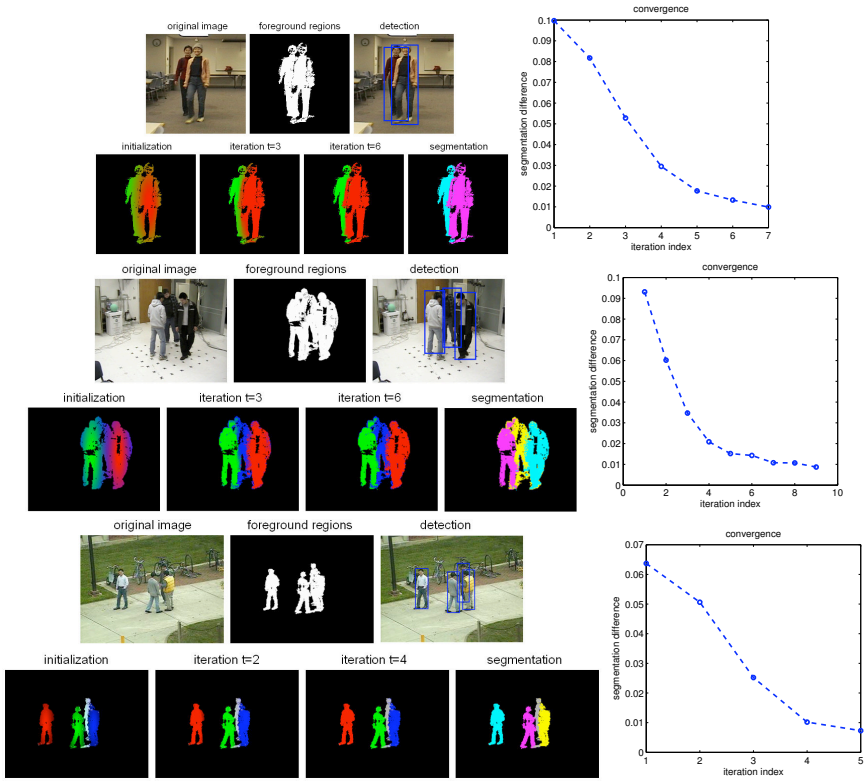

Fig. 2. Examples of the detection and segmentation process with corresponding convergence graphs. The vertical axis of the convergence graph shows the absolute segmentation difference between two consecutive iterations given by Equation 3.

Since we sample test pixels only from the appearance model b, $\hat{p}^b$ is evaluated by its own samples and $\hat{p}^b$ is guaranteed to be equal or larger than $\hat{p}^a$ for any samples $\mathbf{y}_k$. This ensures that the distance $D_{KL}(\hat{p}^b||\hat{p}^a) \geq 0$, where the equality holds if and only if two density models are identical. The Kullback-Leiber distance is a non-symmetric measure in that $D_{KL}(\hat{p}^b||\hat{p}^a) \neq D_{KL}(\hat{p}^a||\hat{p}^b)$. For obtaining a symmetric similarity measure between the two appearance models, we define the distance of the two appearances as follows: $Dist(\hat{p}^b, \hat{p}^a) = \min(D_{KL}(\hat{p}^b||\hat{p}^a), D_{KL}(\hat{p}^a||\hat{p}^b))$. It is reasonable to choose the minimum as the distance measure since it can preserve the balance between (full-full), (full-partial), (partial-partial) appearance matching, while the symmetrized distance $D_{KL}(\hat{p}^b||\hat{p}^a) + D_{KL}(\hat{p}^a||\hat{p}^b))$ would only be effective for (full-full) appearance matching and does not compensate for occlusion.

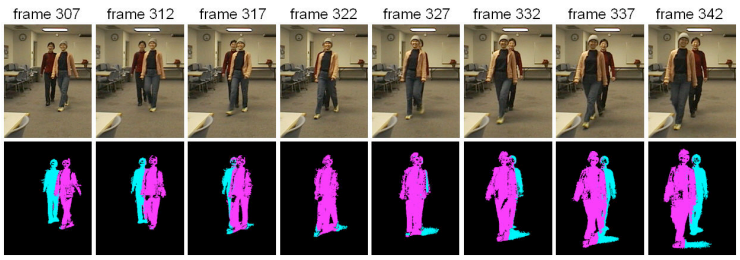

Fig. 3. Experiments on different degrees of occlusion between two people.

5 Experimental Results and Analysis

Fig. 2 shows examples of the human segmentation process for small human groups. The results show that our approach can generate accurate pixel-wise segmentation of foreground regions when people are in standing or walking poses. Also, the convergence graphs show that our segmentation algorithm converges to a stable solution in less than 10 iterations and gives accurate segmentation of foreground regions for images with discriminating color structures of different humans. The cases of falling into local minimum with inaccurate segmentation is mostly due to the color ambiguity between different foreground objects or misclassification of shadows as foreground. Some inaccurate segmentation results can be found in human heads and feet in Fig. 2 and Fig. 3, and can be reduced by incorporating human pose models as in [17].

We also evaluated the segmentation performance with respect to the degree of occlusion. Fig. 3 shows the segmentation results given images with varying degrees of occlusion when two people walk across each other in an indoor environment. Note that the degree of occlusion does not significantly affect the segmentation accuracy as long as reasonably accurate detections can be achieved.

Finally, we quantitatively evaluate our segmentation and appearance modeling approach to appearance matching under occlusion. We choose three frames from a test video sequence (containing two people in the scene) and perform segmentation for each of them. Then, the generated segmentations are used to estimate partial or full human appearance models as shown in Fig. 4. We evaluate the two-way Kullback-Leiber distances and the symmetric distance for each pair of appearances and represent them as affinity matrices shown in Fig. 4. The elements of the affinity matrices quantitatively reflect the accuracy of matching. We also conducted matching experiments using different spatial-color space combinations, 3D (r, g, s) color space, 4D (x, r, g, s) space, 4D (y, r, g, s) space, and 5D (x, y, r, g, s) space. The affinity matrices show that 3D (r, g, s) color space and 4D (y, r, g, s) space produce much better matching results than the other two. This is because color variation is more sensitive in the horizontal direction than in the vertical direction. The color-only feature space obtains good matching performance for this example because the color distributions are

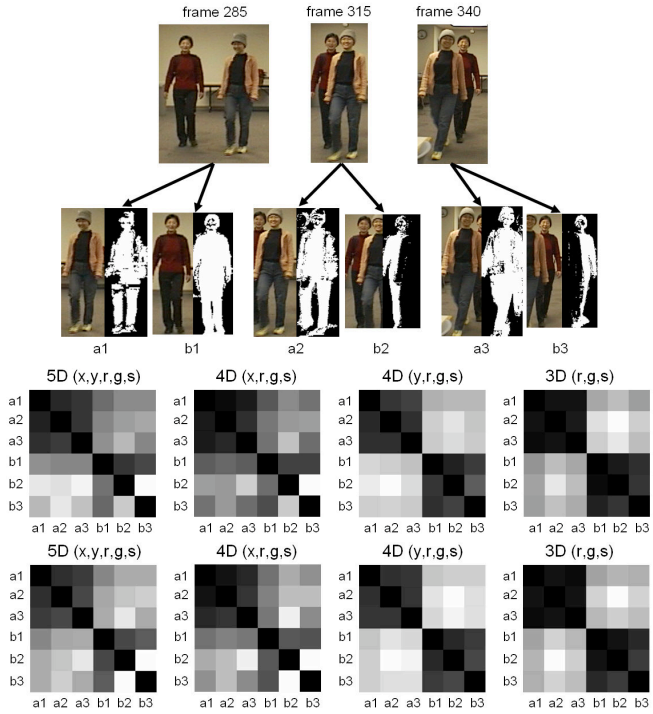

Fig. 4. Experiments on appearance matching. Top: appearance models used for matching experiments, Middle: two-way Kullback-Leiber distances, Bottom: symmetric distances.

significantly different between appearances 1 and 2. But, in reality, there are often cases in which two different appearances have similar color distributions with completely different spatial layouts. On the other hand, 4D (y, r, g, s) joint spatial-color feature space (color distribution as a function of the normalized human height) enforces spatial constraints on color distributions, hence it has much more discriminative power.

6 Conclusion

We proposed a two stage foreground segmentation approach by combining human detection and iterative foreground segmentation. The KDE-EM framework is modified and applied to segmentation of groups into individuals. The advantage of the proposed approach lies in simultaneously segmenting people and building appearance models. This is useful when matching and recognizing people when only occluded frames can be used for training. Our future work includes the application of the proposed approach to human tracking and recognition across cameras.

Acknowledgement

This research was funded in part by the U.S. Government VACE program.

References

1. Zhao, T., Nevatia, R.: Tracking Multiple Humans in Crowded Environment. In: CVPR (2004)
2. Smith, K., Perez, D. G., Odobez, J. M.: Using Particles to Track Varying Numbers of Interacting People. In: CVPR (2005)
3. Rittscher, J., Tu, P. H., Krahnstoever, N.: Simultaneous Estimation of Segmentation and Shape. In: CVPR (2005)
4. Elgammal, A. M., Davis, L. S.: Probabilistic Framework for Segmenting People Under Occlusion. In: ICCV (2001)
5. Mittal, A., Davis, L. S.: M2Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene. International Journal of Computer Vision (IJCV) 51(3) (2003) 189-203
6. Fleuret, F., Lengagne, R., Fua, P.: Fixed Point Probability Field for Complex Occlusion Handling. In: ICCV (2005)
7. Khan, S., Shah, M.: A Multiview Approach to Tracking People in Crowded Scenes using a Planar Homography Constraint. In: ECCV (2006)
8. Kim, K., Davis, L. S.: Multi-Camera Tracking and Segmentation of Occluded People on Ground Plane using Search-Guided Particle Filtering. In: ECCV (2006)
9. Tao, H., Sawhney, H., Kumar, R.: A Sampling Algorithm for Detecting and Tracking Multiple Objects. In: ICCV Workshop on Vision Algorithms (1999)
10. Isard, M., MacCormick, J.: BraMBLe: A Bayesian Multiple-Blob Tracker. In: ICCV (2001)
11. Scott, D. W.: Multivariate Density Estimation. Wiley Interscience (1992)
12. Elgammal, A. M., Davis, L. S.: Probabilistic Tracking in Joint Feature-Spatial Spaces. In: CVPR (2003)
13. Zhao, L., Davis, L. S.: Iterative Figure-Ground Discrimination. In: ICPR (2004)
14. Zhao, L., Davis, L. S.: Segmentation and Appearance Model Building from An Image Sequence. In: ICIP (2005)
15. Wang, H., Suter, D.: Tracking and Segmenting People with Occlusions by A Simple Consensus based Method. In: ICIP (2005)
16. Lin, Z., Davis, L. S., Doermann, D., DeMenthon D.: Hierarchical Part-Template Matching for Human Detection and Segmentation. In: ICCV (2007)
17. Lin, Z., Davis, L. S., Doermann, D., DeMenthon D.: An Interactive Approach to Pose-Assisted and Appearance-based Segmentation of Humans. In: ICCV Workshop on Interactive Computer Vision (2007)
18. Yu, Y., Harwood, D., Yoon, K., Davis, L. S.: Human Appearance Modeling for Matching across Video Sequences. Special Issue on Machine Vision Applications (2007)