Monocular Video Foreground/Background Segmentation by Tracking Spatial-Color Gaussian Mixture Models

Ting Yu*
GE Global Research
Niskayuna, NY 12309

Cha Zhang, Michael Cohen, Yong Rui
Microsoft Research
Redmond, WA 98052

Ying Wu
ECE, Northwestern University
Evanston, IL 60208

Abstract

This paper presents a new approach to segmenting monocular videos captured by static or hand-held cameras filming large moving non-rigid foreground objects. The foreground and background objects are modeled using spatial-color Gaussian mixture models (SCGMM), and segmented using the graph cut algorithm, which minimizes a Markov random field energy function containing the SCGMM models. In view of the existence of a modeling gap between the available SCGMMs and segmentation task of a new frame, one major contribution of our paper is the introduction of a novel foreground/background SCGMM joint tracking algorithm to bridge this space, which greatly improves the segmentation performance in case of complex or rapid motion. Specifically, we propose to combine the two SCGMMs into a generative model of the whole image, and maximize the joint data likelihood using a constrained Expectation-Maximization (EM) algorithm. The effectiveness of the proposed algorithm is demonstrated on a variety of sequences.

1 Introduction

Segmenting foreground objects from the background in videos is of great interest in many applications. In video conferencing, once the foreground and background are separated, the background can be replaced by another image, which then beautifies the video and protects the user privacy. The extracted foreground objects can be compressed to facilitate efficient transmission using object-based video coding. As an advanced video editing tool, segmentation also allows people to combine multiple objects from different videos and create new and artistic results.

In this paper, we study the foreground/background segmentation problem for monocular videos. In particular, we are interested in videos captured by static or hand-held cameras filming large moving non-rigid foreground objects. For instance, the foreground can be the head and shoulder of a talking person, or a dancing character. We relax the static background restriction, and assumes that the camera can be shaking to some extent and there can be other moving objects in the background. The main challenges in such kind of sequences are:

- Portions of the foreground and background objects may share similar color patterns.
- The sizes of foreground objects for such applications are relatively large, hence substantial occlusions may frequently occur between the foreground and background objects.
- The objects being segmented can have non-rigid appearance. The foreground objects may also demonstrate complex and rapid motions, which can consequently fail the flow computation step of many existing approaches.
- The existence of other uninteresting but moving background objects may also cause additional confusions to the segmentation algorithm if they are not correctly modeled.

Compared with the work in [12] that utilizes the depth information reconstructed from a stereo camera pair, monocular video foreground/background segmentation is much less constrained. Additional assumptions are often made to limit the solution space. For instance, the background scene can be assumed as static and known a prior, which converts the segmentation problem into a background modeling and subtraction problem. Existing solutions include pixel level modeling using Gaussian distributions [24], mixture of Gaussians [19], non-parametric kernel density estimators [7, 15] and three state HMM [16]. A separate region-level or even object-level model could be added to further improve the background modeling quality [9, 23, 20]. Nevertheless, video segmentation based on background modeling may still be confused by moving background objects or motionless foreground objects.

Another popular assumption is that the foreground and background objects have different motion patterns. Research in this vein, termed as layer-based motion segmentation (LBMS), received tremendous interests in the past decades [1, 22, 10, 25, 13]. The general objective is to automatically extract the motion coherent regions. Primarily focusing on the general motion segmentation problem,

*Work performed while at Microsoft Research.

existing approaches in LBMS were either computationally expensive requiring off-line learning and a batch processing of the whole sequences [1, 25, 10, 13], or tended to generate many over-segmented regions, thus hard to form semantically meaningful objects [22]. In [5], a discriminative model was learned to efficiently separate the motion pixels from the stasis using pixel spatio-temporal derivatives. However, such a generic motion classifier may be challenged when camera is shaking or background objects are moving.

In this paper, we propose to model both the foreground and background objects using SCGMM models. These models are built into a Markov random field (MRF) energy function that is then efficiently minimized by the graph cut algorithm [3], leading to a binary segmentation of the video frames.

While it is well-known that SCGMM has better discriminative power than color-only GMM, integrating it with the graph cut based video foreground/background segmentation algorithm is nontrivial. For scenes with complex and rapid foreground/background motions, the SCGMM learned from previous frames are not suitable for the segmentation task of the current frame due to large variations of the spatial components. The major contribution of this paper is the introduction of a foreground/background SCGMM joint tracking algorithm, which can reliably propagate the SCGMM models over frames. To achieve this, before the segmentation of the current frame is carried out, we first combine the foreground and background SCGMMs learned from previous frames into a single generative model to depict the whole image, then adopt an EM algorithm to update the model for the new frame under the constraint that the color factors of the objects remain unchanged. The updated whole image SCGMM is then split back into two SCGMMs and used for segmentation. The proposed method can not only handle complex and rapid motions of the foreground/background objects, but also help resolve occlusion/deocclusion issues caused by their motions. The effectiveness of the algorithm is verified by a variety of challenging sequences.

The proposed method is different from recent approaches that also make use of spatial and color models [11, 6, 18]. For example, [6, 18] proposed to model the object spatial-color features with kernel density estimation. An assumption they made is that the motions between subsequent frames are small, hence the mixture models in the previous frame can be directly applied to the segmentation of the current frame. Our study, however, shows that such a simplified assumption is not valid for many real world sequences. It is our main contribution that we explicitly introduce the foreground/background SCGMM joint tracking step to fill in this gap. Our experimental results also strongly suggested that such a tracking step is necessary for successful segmentation of the sequences.

The paper is organized as follows. The formulation of the MRF energy function and its minimization through graph cut is described in Section 2. The foreground/background joint SCGMM tracking algorithm is presented in Section 3. One possible option of automatic system initialization for teleconferencing applications is discussed in Section 4. Experimental results and conclusions are given in Section 5 and 6, respectively.

2 Energy Minimization Formulation

We propose to solve the foreground/background segmentation problem from video using energy minimization. At any time instant t, let the feature vectors extracted from the video pixels be $\mathbf{z}_{i,t}, i = 1, \cdots, N$, where N is the number of pixels in each frame. Denote the unknown label of each pixel as $f_{i,t}, i = 1, \cdots, N$, where $f_{i,t}$ is a binary variable, i.e., 0 and 1 in our case with $f_{i,t} = 1$ representing pixel i labeled as foreground, and $f_{i,t} = 0$ as background [2]. In the following discussions, we may ignore subscript t when it causes no confusion.

An energy-based objective function can be formulated over the unknown labeling variables of every pixel, $f_i, i = 1, \ldots, N$, in the form of a first-order Markov random field (MRF) energy function:

$$\begin{aligned} E(f) &= E_{data}(f) + \lambda E_{smooth}(f) \\ &= \sum_{p \in \mathcal{P}} D_p(f_p) + \lambda \sum_{\{p,q\} \in \mathcal{N}} V_{p,q}(f_p, f_q), \end{aligned} \quad (1)$$

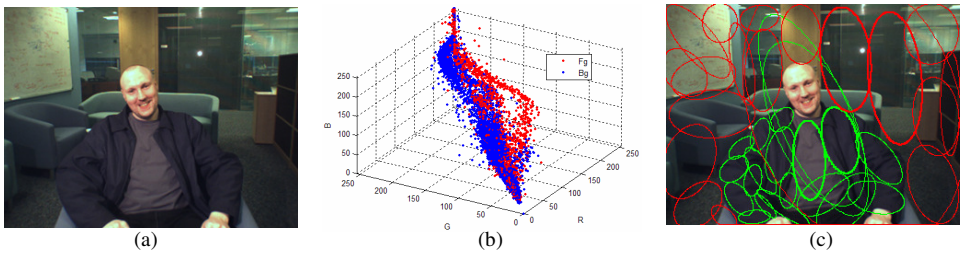
where $\mathcal{N}$ denotes the set of 8-connected pair-wise neighboring pixels, $\mathcal{P}$ is the set of pixels in each image. The role of λ is to balance the data $D_p(f_p)$ and smooth cost $V_{p,q}(f_p, f_q)$. The above energy function can be efficiently minimized by a two-way graph cut algorithm [3], where the two terminal nodes represent foreground and background labels respectively.

We model the pair-wise smoothness energy term $E_{smooth}(f)$ as:

$$\begin{aligned} E_{smooth}(f) &= \sum_{\{p,q\} \in \mathcal{N}} V_{p,q}(f_p, f_q) \\ &= \sum_{\{p,q\} \in \mathcal{N}} \frac{1}{d(p,q)} e^{-\frac{(I_p - I_q)^2}{2\sigma^2}}. \end{aligned} \quad (2)$$

where I_p denotes the intensity of pixel p, σ is the average intensity difference between neighboring pixels in the image, and $d(p, q)$ is the distance between two pixels p and q. This smoothness constraint penalizes the labeling discontinuities of neighboring pixels if they have similar pixel intensities. It favors the segmentation boundary along regions where strong edges are detected.

The data energy term $E_{data}(f)$ evaluates the likelihood of each pixel belonging to the foreground or background.



(a) (b) (c)

Figure 1: Color distribution of the scene *JM*. (a) First frame of *JM*. (b) Foreground/Background color confusions in 3-dimensional RGB space, where the red points depict the foreground pixels and the blue points show the background pixels. There are significant overlaps between foreground/background pixels. (c) The spatial-color GMM model of the image. Each ellipse represent a Gaussian component. The green ellipses are the foreground components; the red ellipses are the background components. They are spatially apart so there is much less confusion.

In previous approaches for image segmentation [2, 17, 12, 20, 5], this term is often computed using Gaussian mixture models (GMM) in the RGB color space. Figure 1(b) shows the color distributions of foreground/background objects for the first frame of the test sequence *JM*. It can be seen that the foreground and background pixels have significant overlap in the RGB space, which consequently leads to severe confusion for the data energy term. In this paper, we resort to a SCGMM model to overcome this problem.

We take a five dimensional feature vector to describe each pixel, i.e., $\mathbf{z}_i = (x, y, r, g, b)$, representing the pixel's spatial information, (x, y) coordinates, and color information, (r, g, b) color values. A five dimensional SCGMM model is obtained for each video frame (details in Section 3). The likelihood of a pixel belonging to the foreground or background can be written as:

$$p(\mathbf{z}|l) = \sum_{k=1}^{K_l} p_{l,k} G(\mathbf{z}; \mu_{l,k}, \mathbf{\Sigma}_{l,k}) \tag{3}$$

where $l \in \{fg, bg\}$, representing foreground or background; $p_{l,k}$ is the prior of the k_{th} Gaussian component in the mixture model, and $G(\mathbf{z}; \mu_{l,k}, \mathbf{\Sigma}_{l,k})$ is the k_{th} Gaussian component as:

$$G(\mathbf{z}; \mu_{l,k}, \mathbf{\Sigma}_{l,k}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\mathbf{\Sigma}_{l,k}|^{\frac{1}{2}}} e^{-\frac{(\mathbf{z}-\mu_{l,k})^T \mathbf{\Sigma}_{l,k}^{-1} (\mathbf{z}-\mu_{l,k})}{2}}, \tag{4}$$

where $d = 5$ is the dimension of the GMM models.

We further assume that the spatial and color components of the GMM models are decoupled, i.e., the covariance matrix of each Gaussian component takes the block diagonal form, $\mathbf{\Sigma}_{l,k} = \begin{pmatrix} \mathbf{\Sigma}_{l,k,s} & 0 \\ 0 & \mathbf{\Sigma}_{l,k,c} \end{pmatrix}$, where s and c stand for the spatial and color features respectively. With such decomposition, each GMM Gaussian component has the following factorized form:

$$G(\mathbf{z}; \mu_{l,k}, \mathbf{\Sigma}_{l,k}) = G(\mathbf{z}_s; \mu_{l,k,s}, \mathbf{\Sigma}_{l,k,s}) G(\mathbf{z}_c; \mu_{l,k,c}, \mathbf{\Sigma}_{l,k,c}). \tag{5}$$

The SCGMM models for the first frame of *JM* is shown in Figure 1(c). Each ellipse represents a Gaussian component of the foreground (green) or background (red). The thickness of each ellipse indicates the component weight $p_{l,k}$. It is clear that the foreground and background SCGMMs are spatially apart from each other, thus rendering more discriminative power than the color-only GMMs.

Given the SCGMM models, the data cost $E_{data}(f)$ is defined as:

$$E_{data}(f) = \sum_{p \in \mathcal{P}} D_p(f_p) = \sum_{p \in \mathcal{P}} -\log p(\mathbf{z}_p|f_p), \tag{6}$$

where $p(\mathbf{z}_p|f_p)$ is computed using Equation 3.

The above SCGMM model can be extended to more sophisticated models. For example, we can form a global-local mixture distribution for the background model by combining the existing spatial-color components with a localized per-pixel background model [12, 20]. This may be more suitable to represent the multi-modality nature of background pixels in outdoor surveillance video. However, since the primary use of our system is for indoor office environment, we chose not to explore further along this direction.

As we acknowledged previously, recent approaches also employ the idea of spatial augmented GMM models for video segmentation [24, 11, 6, 18]. However, most existing approaches hold a small object motion assumption, hence conclude that the mixture models in the previous frame can be directly applied to the segmentation of the current frame. Our comprehensive study shows that this assumption is not valid for many real world sequences. As we observed, when the objects of interest undergo large motions, the spatial-color mixture models obtained only from previous frame can bring a significant bias towards the segmentation and dramatically degrade the performance. Hence, we propose to address this problem using a SCGMM joint tracking algorithm in the next section, which can successfully fill in such gap between model and data.

3 SCGMM Joint Tracking

3.1 Problem Statement

Suppose two SCGMM models are learned during the system initialization period (see Section 4) using the popular EM algorithm, which maximizes the data likelihood of each segment:

$$\theta^*_{l,0} \stackrel{\text{def}}{=} \{p^*_{l,k,0}, \mu^*_{l,k,0}, \mathbf{\Sigma}^*_{l,k,0}\} = \underset{p^*_{l,k,0}, \mu^*_{l,k,0}, \mathbf{\Sigma}^*_{l,k,0}}{\arg\max} \prod_{\mathbf{z}_l \in I_0} [\sum_{k=1}^{K_l} p_{l,k} G(\mathbf{z}_l; \mu_{l,k}, \mathbf{\Sigma}_{l,k})] \tag{7}$$

where $l \in \{fg, bg\}$; $\mathbf{z}_l$ are the features of the pixels having label l; I_0 denotes the initialization frame. The problem with video segmentation is how to propagate these SCGMM models over the rest of the sequence, since both the foreground and background objects can be constantly moving.

We consider the following general problem. Suppose at time instant $t-1$, the parameters of the foreground and background SCGMM models have been learned as $\theta_{l,t-1} = (p_{l,k,t-1}, \mu_{l,k,t-1}, \mathbf{\Sigma}_{l,k,t-1}), k = 1, \cdots, K_l$. Given a newly coming frame I_t, the goal is to obtain a foreground/background segmentation for I_t, and update the SCGMM models $\theta_{l,t} = (p_{l,k,t}, \mu_{l,k,t}, \mathbf{\Sigma}_{l,k,t}), k = 1, \cdots, K_l$. At the first glance, this problem appears to be a deadlock, because in order to obtain a good segmentation, we need an accurate SCGMM model for the *current* frame; and in order to get an accurate SCGMM model for the current frame, we need a good foreground/background segmentation. Below we present a SCGMM joint tracking algorithm to break this deadlock.

3.2 SCGMM Joint Tracking

We look for ways to obtain an approximate SCGMM model for the current frame before the graph cut segmentation. Inspired by color-based object trackers such as [4], we assume that from time $t-1$ to t, the colors of the foreground and background objects do not change. Hence, the color parts of the SCGMM models remain identical:

$$G(\mathbf{z}_{c,t}; \mu_{l,k,c,t}, \mathbf{\Sigma}_{l,k,c,t}) = G(\mathbf{z}_{c,t-1}; \mu_{l,k,c,t-1}, \mathbf{\Sigma}_{l,k,c,t-1}), \tag{8}$$

where c denotes the color dimension and $k = 1, \cdots, K_l$. The problem then becomes how to formulate an updating scheme for the spatial parts $G(\mathbf{z}_{s,t}; \mu_{l,k,s,t}, \mathbf{\Sigma}_{l,k,s,t})$ given the new input image I_t.

Since we do not have a foreground/background segmentation on I_t at this moment, we first form a global SCGMM model of the whole image by combining the foreground and background SCGMM models of the previous frame, with the corresponding weights equal to the relative coverage sizes of foreground and background regions in the previous frame, i.e., we define:

$$\theta^0_{I,t} \stackrel{\text{def}}{=} \{\theta_{fg,t-1}, \theta_{bg,t-1}, \gamma_{fg,t-1}, \gamma_{bg,t-1}\}, \tag{9}$$

where superscript 0 indicates that the parameter set is serving as the initialization value for the later update. $\gamma_{fg,t-1}$ and $\gamma_{bg,t-1}$ represent the weights or coverage areas of the foreground and background regions in the previous segmented frame, and they satisfy $\gamma_{fg,t-1} + \gamma_{bg,t-1} = 1$.

Denote $K_I = K_{fg} + K_{bg}$ as the number of Gaussian components in the combined image level SCGMM model, where we assume the first K_{fg} Gaussian components are from the foreground SCGMM, and the last K_{bg} Gaussian components are from the background SCGMM. The image SCGMM model can be written as:

$$\begin{aligned} p(\mathbf{z}_t|\theta^0_{I,t}) &= \gamma_{fg,t-1} p(\mathbf{z}_t|\theta_{fg,t-1}) + \gamma_{bg,t-1} p(\mathbf{z}_t|\theta_{bg,t-1}) \\ &= \sum_{k=1}^{K_I} p^0_{k,t} G(\mathbf{z}_{s,t}; \mu^0_{k,s,t}, \mathbf{\Sigma}^0_{k,s,t}) G(\mathbf{z}_{c,t}; \mu_{k,c,t}, \mathbf{\Sigma}_{k,c,t}). \end{aligned} \tag{10}$$

Note the second Gaussian term over the color dimension is defined in Equation 8 and remains fixed at this moment. The Gaussian component weights $p^0_{k,t}, k = 1, \ldots, K_I$ are different from their original values in their individual foreground or background SCGMMs due to the multiplications of $\gamma_{fg,t-1}, \gamma_{bg,t-1}$.

Given the pixels in the current frame I_t, our objective is to obtain an updated parameter set $\{p^*_{k,t}, \mu^*_{k,s,t}, \mathbf{\Sigma}^*_{k,s,t}\}$ over the spatial domain, which maximizes the joint data likelihood of the whole image, i.e.,

$$\{p^*_{k,t}, \mu^*_{k,s,t}, \mathbf{\Sigma}^*_{k,s,t}\} = \underset{p^*_{k,t}, \mu^*_{k,s,t}, \mathbf{\Sigma}^*_{k,s,t}}{\arg\max} \prod_{\mathbf{z}_t \in I_t} p(\mathbf{z}_t|\theta_{I,t}) \tag{11}$$

for all $k = 1, \ldots, K_I$. The EM algorithm is adopted here to iteratively update the model parameters from their initial values $\theta^0_{I,t}$. However, as can be seen in Eq. 11, unlike the traditional EM algorithm, where all model parameters are simultaneously updated, we choose to only update the spatial parameters of the SCGMM models in this phase, and keep the color parameters unchanged. This can be implemented by constraining the color mean and variance to be fixed to their corresponding values in the previous frame (Equation 8).

Such a restricted EM algorithm is shown in Figure 2. In the E-step, we calculate the posteriori of the pixels belonging to each Gaussian component, and in the M-step, the mean and variance of each Gaussian component in spatial domain are refined based on the updated posteriori probability of pixel assignment from E-step. In the statistics literature, such a variant of EM algorithm aiming to maximizing the conditional data likelihood (over spatial vari-

At each time instant t, we perform the following EM algorithm:

1. E-step, calculate the Gaussian component assignment probability for each pixel $\mathbf{z}$:

$$p^{(i)}(k|\mathbf{z}) = \frac{p_k^{(i)} G(\mathbf{z}_s; \mu_{k,s}^{(i)}, \mathbf{\Sigma}_{k,s}^{(i)}) G(\mathbf{z}_c; \mu_{k,c}, \mathbf{\Sigma}_{k,c})}{\sum_{k=1}^{K_I} p_k^{(i)} G(\mathbf{z}_s; \mu_{k,s}^{(i)}, \mathbf{\Sigma}_{k,s}^{(i)}) G(\mathbf{z}_c; \mu_{k,c}, \mathbf{\Sigma}_{k,c})}.$$

2. M-step, update the spatial mean and variance, and the weight of each Gaussian component as:

$$\mu_{k,s}^{(i+1)} = \frac{\sum_{\mathbf{z} \in I_t} p^{(i)}(k|\mathbf{z}) \mathbf{z}_s}{\sum_{\mathbf{z} \in I_t} p^{(i)}(k|\mathbf{z})}.$$

$$\mathbf{\Sigma}_{k,s}^{(i+1)} = \frac{\sum_{\mathbf{z} \in I_t} p^{(i)}(k|\mathbf{z}) (\mathbf{z}_s - \mu_{k,s}^{(i+1)})(\mathbf{z}_s - \mu_{k,s}^{(i+1)})^T}{\sum_{\mathbf{z} \in I_t} p^{(i)}(k|\mathbf{z})}$$

$$p_k^{(i+1)} = \frac{\sum_{\mathbf{z} \in I_t} p^{(i)}(k|\mathbf{z})}{\sum_{k=1}^{K_I} \sum_{\mathbf{z} \in I_t} p^{(i)}(k|\mathbf{z})}$$

Figure 2: The EM algorithm for foreground/background joint tracking. Subscript t is ignored in the above equations.

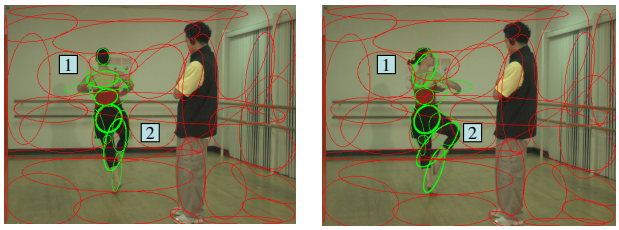


Figure 3: Foreground/background joint SCGMM tracking via EM. Ellipses in green are foreground components, and ellipses in red are background components. Note the expansion of background component 1 and the compression of component 2 while the foreground person rotates.

ables) is called an Expectation Conditional Maximization algorithm [14].

The above algorithm shares many common characteristics with the gradient-based color tracking algorithms in literature such as mean-shift [4], hence we name it as *SCGMM joint tracking*. For instance, the spatial part of the SCGMM resembles the spatial kernel used in mean-shift. Both approaches are gradient-based, which moves the component/object bounding box towards directions where there is more color similarity. Tao et al. also proposed an EM based object tracking algorithm in [21], though each object layer was modeled using a single Gaussian. Compared with the existing methods, the proposed SCGMM joint tracking has a number of unique features:

1. Unlike many existing algorithms that only focus on tracking the foreground objects, the proposed algorithm combines the foreground and background SCGMMs into a unified model and tracks both simultaneously. The tracking is performed through maximizing the overall data likelihood of the whole image. Hence the foreground and background SCGMMs can collaborate with each other and adapt better to the change of the whole image.

2. Since in the E-step of Figure 2, the pixels are assigned to different Gaussian components based on their likelihoods, the foreground and background Gaussian components are actually also competing with each other to grab similar pixels. This partially solves the occlusion/deocclusion problem in video segmentation. As shown in Figure 3, when the foreground object rotates and moves to the right, the background component 1 on the left is expanded, and the background component 2 on the right is compressed.

3. Objects with very complex colors or shapes can be tracked, thanks to the descriptive capability of the SCGMM. The *handset* sequence in Section 5 was tracked using multiple collaborative kernel tracking in [8], which requires the knowledge of the structure of the handset before hand. It is successfully tracked and segmented with our technique without any special treatment.

4. The SCGMM can track highly deformable non-rigid objects. The *Ballet* sequence in Section 5 shows the tracking and segmentation of a human dancer. Although the foreground SCGMM does not describe each semantic body part very accurately, the part-based nature of the SCGMM model renders the proposed algorithm capable of tracking and segmenting such highly articulated objects.

3.3 Segmentation and Post-Updating

After the SCGMM joint tracking, the image SCGMM model is split back into two models, one describing the foreground, the other describing the background. Components belonging to the foreground before tracking are placed in the foreground SCGMM, and components belonging to the background before tracking are placed in the background SCGMM. The two SCGMM models are then used to perform graph cut segmentation, as described in Section 2.

The segmentation results can be used for a post-updating of the SCGMM models, because now we can train the foreground and background SCGMMs separately with the segmented pixels, which often provides better discriminative power for segmenting future frames. The process is similar to what we did for the model initialization (Equation 7), except that we will use the tracked SCGMM models as the

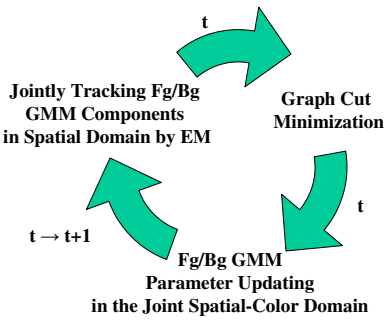


Figure 4: The Iterative Circle of Foreground/Background Segmentation for One Frame.

initialization for the optimization process, which often results in a faster convergence.

The segmentation results are not always perfect. If we believe that the foreground and background colors stay the same across the sequence, we can perform a constrained update on the two models. That is, we use Equation 11 on the foreground or background region to update the SCGMM models, forcing the color means and variances to be constant. Our experiments show that this approach will often help the whole application to recover from segmentation errors. We also tried to update both the spatial and color components after segmentation. We found it a risky approach because it is very volunable to error propagation when there are segmentation errors in a certain frame.

To summarize, we iterate the tracking-segmentation-updating process as shown in Figure 4. In principle, for each frame the circle can be run several times until convergence. In practice we found one iteration is sufficient for all the sequences we tested.

4 Automatic Foreground Extraction for Video Conferencing

There are many options for our system to obtain a model initialization for the foreground/background objects. In particular, for real-time applications such as video conferencing, one possibility is to initialize with motion information, as the foreground user can be asked to present large motions during the system initialization period. Here we briefly describe another automatic initialization approach by assuming we know that the to-be-segmented foreground is the head and shoulder of a person.

The first step is to apply a face detector [26] on the video frame, as shown in Figure 5(a). Based on the detection results, we assume certain regions to be definite foreground (shown in white in Figure 5(b)) and background (shown in black). These play a similar role to the strokes drawn by the users in interactive image segmentation [2, 17]. For instance, the middle of the detected face rectangle are guaranteed to be foreground; the shoulders are likely to be a slightly expanded region below the detected face. Similarly, the areas to the left, right and above all expanded face rectangles are assumed to be background (the black areas in Figure 5(b)). We train two SCGMM models from the pixels covered by the definite foreground and background regions, and perform a graph cut segmentation. This gives us Figure 5(c). The segmentation is not perfect, however they are sufficient to initialize the models for the whole sequence. In Section 5 we will show the segmentation results of the two sequences in Figure 5 with automatic initialization.

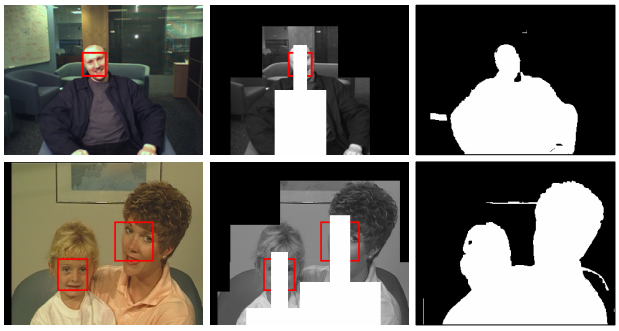


Figure 5: Automatic initialization of the segmentation in the first frame. (a) Face detection results. (b) We assume certain regions to be definite foreground and background. White indicates foreground, black indicates background. (c) Image segmentation result using graph cut.

5 Experimental Results

The proposed video foreground/background segmentation algorithm is applied to a variety of sequences to verify the performance. In the following, we describe the challenges in each sequence, and demonstrate the effectiveness of our algorithm as shown in Figure 6.

Mother and Daughter: This scene is relatively simple, as there is no significant movement in the foreground or background. We initialize the segmentation in the first frame using Figure 5(c). The algorithm works very well across the whole sequence.

Foreman: *Foreman* is captured with a hand-held camera. Hence there are consistent motions for both foreground and background, which may cause trouble for both pixel-wise background subtraction and layer-based motion segmentation approaches. Our tracking algorithm is not affected by such motions. One difficulty we face in this sequence is that from frame #95 to #108, the left shoulder of the person is completely out of the field of view of the camera. This cannot be immediately recovered, as shown in Figure 6, frame #135. The SCGMM joint tracking in this case plays an important role to expand the coverage of the foreground region

Figure 6: Segmentation results with the proposed algorithm.

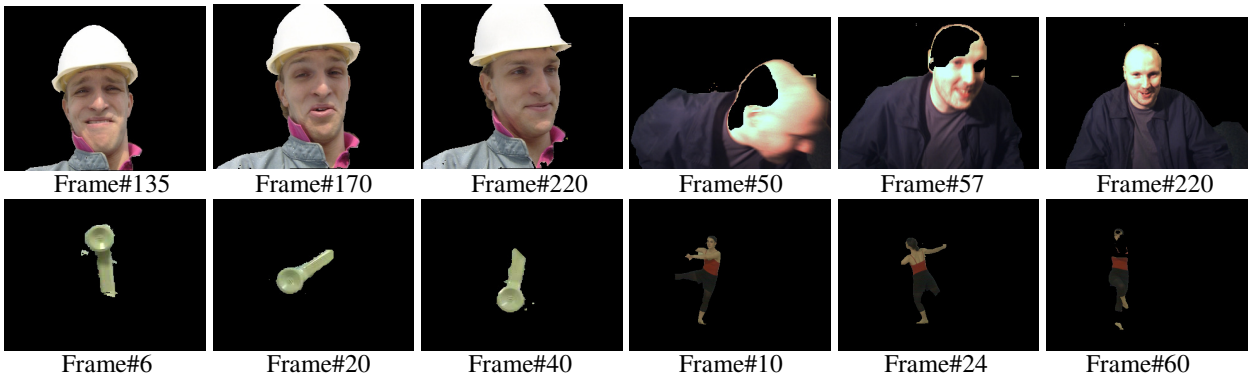


Figure 7: Segmentation results without joint SCGMM tracking. We do perform post-updating as described in Section 3.3. This way of applying SCGMM is similar to many existing algorithms [24, 6].

(frame #170). In contrast, as shown in Figure 7, segmentation without this intermediate tracking step was not able to recover the missing part even in frame #220.

JM: *JM* is a very typical video conferencing sequence. We initialize the first frame automatically using the method introduced in Section 4. The main challenge of this sequence is between frame #42 and #61, where the person makes some dramatic movements. The proposed method handles this period very well. On the contrary, segmentation without tracking produces unacceptable results (Figure 7).

Handset: This is a challenging sequence due to the fast rotation of the handset. The video was heavily compressed,

thus the segmentation boundaries are noisy. Compared with the no-tracking segmentation results in Figure 7, which is a complete failure, the outputs of the proposed algorithm are very satisfactory.

Ballet: *Ballet* is probably the most difficult sequence we tested due to the fast motion of the arms and legs of the dancer. The results generated by the proposed algorithm is surprisingly good, in particular in the level of details of the dancer's hands. Though we do have a failure around frame #24, where one foot is missing due to the complete occlusion of this foot in frame #15, such a failure is also successfully recovered in later frames. Figure 7 shows the segmentation results without the proposed tracking step, and the performance is very poor.

6 Conclusions

We have proposed a novel joint tracking algorithm for spatial-color Gaussian mixture models (SCGMM) in monocular video foreground/background segmentation. The basic idea is to combine the foreground and background SCGMM models into a generative model of the whole image, and use a variant of EM algorithm to update the model under the equality constraint that the color factors of the model do not change. We show that this algorithm improves the segmentation results significantly on a number of challenging sequences.

References

[1] E. H. Adelson and J. Y. A. Wang. Representing moving images with layers. *In IEEE Trans. on Image Processing*, 1994.

[2] Y. Boykov and M. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. In *Proc. IEEE Int'l Conf. on Computer Vision (ICCV)*, volume I, pages 105–112, 2001.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *In IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 23(11):1222–1239, 2001.

[4] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 142–149, 2000.

[5] A. Criminisi, G. Cross, A. Blake, and V. Kolmogorov. Bilayer segmentation of live video. In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[6] A. Elgammal and L. S. Davis. Probabilistic framework for segmenting people under occlusion. In *Proc. IEEE Int'l Conf. on Computer Vision (ICCV)*, Vancouver, Canada, July 2001.

[7] A. Elgammal, D. Harwood, and L. S. Davis. Non-parametric model for background subtraction. In *Proc. European Conf. on Computer Vision (ECCV)*, Dublin, Ireland, June 2000.

[8] Z. Fan and Y. Wu. Multiple collaborative kernel tracking. In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, 2005.

[9] M. Harville. A framework for high-level feedback to adaptive per-pixel, mixture-of-gaussian background models. In *Proc. European Conf. on Computer Vision (ECCV)*, 2002.

[10] N. Jojic and B. Frey. Learning flexible sprites in video layers. In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, Hawaii, 2001.

[11] S. Khan and M. Shah. Object based segmentation of video using color, motion and spatial information. In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[12] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother. Bilayer segmentation of binocular stereo video. In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume I, San Diego, CA, 2005.

[13] P. Kumar, P. Torr, and A. Zisserman. Learning layered motion segmentations of video. In *Proc. IEEE Int'l Conf. on Computer Vision (ICCV)*, Beijing, China, Oct. 2005.

[14] X. Meng and D. Rubin. Maximum likelihood estimation via the ecm algorithm: A general framework. *In Biometrika*, 80(2), 1993.

[15] A. Mittal and N. Paragios. Motion-based background subtraction using adaptive kernel density estimation. In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, Washington, D. C., June 2004.

[16] J. Rittscher, J. Kato, S. Joga, and A. Blake. A probabilistic background model for tracking. In *Proc. European Conf. on Computer Vision (ECCV)*, 2000.

[17] C. Rother, V. Kolmogorov, and A. Blake. Grabcut - interactive foreground extraction using iterated graph cuts. In *Proc. Siggraph*, 2004.

[18] Y. Sheikh and M. Shah. Bayesian object detection in dynamic scenes. In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, June 2005.

[19] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 246–252, 1999.

[20] J. Sun, W. Zhang, X. Tang, and H. Y. Shum. Background cut. In *Proc. Europ. Conf. on Computer Vision (ECCV)*, Graz, Austria, 2006.

[21] H. Tao, H. Sawhney, and R. Kumar. Object tracking with bayesian estimation of dynamic layer representations. *In IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24:75–89, 2002.

[22] P. H. S. Torr, R. Szeliski, and P. Anandan. An integrated bayesian approach to layer extraction from image sequences. *In IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2001.

[23] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower : Principles and practice of background maintenance. In *Proc. Int'l Conf. on Computer Vision (ICCV)*, 1999.

[24] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland. Pfinder: Real time tracking of the human body. *In IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1997.

[25] J. J. Xiao and M. Shah. Motion layer extraction in the presence of occlusion using graph cut. In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, Washington, D. C., June 2004.

[26] Z. Zhang, M. Li, S. Li, and H. Zhang. Multi-view face detection with floatboost. In *Proc. IEEE Workshop on Applications of Computer Vision*, 2002.