Human Segmentation Algorithm for Real-time Video-call Applications

Seon Heo*, Hyung Il Koo†, Hong Il Kim‡, Nam Ik Cho*
*Department of Electrical and Computer Engineering, Seoul National University, Seoul, Republic of Korea.
E-mail: hsfra111@ispl.snu.ac.kr, nicho@snu.ac.kr
†Division of Electrical and Computer Engineering, Ajou University, Suwon, Republic of Korea.
E-mail: hikoo@ajou.ac.kr
‡Samsung Electronics Co. Ltd., Suwon, Republic of Korea.
E-mail: hongil79.kim@samsung.com

***Abstract*—This paper presents a human region segmentation algorithm for real-time video-call applications. Unlike conventional methods, the segmentation process is automatically initialized and the motion of cameras is not restricted. To be precise, our method is initialized by face detection results and human/background regions are modeled with spatial color Gaussian mixture models (SCGMMs). Based on the SCGMMs, we build a cost function considering spatial and color distributions of pixels, region smoothness, and temporal coherence. Here, the temporal coherence term allows us to have stable segmentation results. The cost function is minimized by the well-known graph-cut algorithm and we update our SCGMM models with the segmentation results. Experimental results have shown that our method yields stable segmentation results with a small amount of computation load.**

I. Introduction

Since many video sequences can be divided into foreground and backgrounds layers and people are usually interested in foreground, foreground/background segmentation enables a lot of interesting applications. Some examples are video analysis, region-of-interest (ROI) based video coding, and video editing. Also, as smartphones and webcams are widely available in daily communications, a real-time human segmentation algorithm has been receiving growing attention for video-call applications such as background substitution. Unlike off-line applications having no strict constraints on processing time and user interactions, we have to consider a number of issues in video-call environments: speed, accuracy, user-interaction, and constraints on camera motions. For instance, many fixed-camera based algorithms are efficient and automatic, however, they cannot be used in video-call applications (using hand-held cameras) due to their hard constraints on camera motions. Interactive methods such as [1], [2] allow us to have accurate segmentation results, however, the requirement on user interaction are impractical in video-call environments.

In order to alleviate these limitations, we develop a new segmentation method without assumptions on camera motions. Our method is based on energy minimization framework [3], [4]: the problem is formulated as a labeling-problem that assigns labels to pixels and we design a cost function so that its minimization yields optimal labels. Since the graph-cut algorithm [4] provides efficient and effective energy minimization results, many researchers focussed on the design of cost functions consisting of data terms and prior terms [2].

For the design of data terms, non-parametric models were used in [5] and pixel-wise Gaussian models were adopted in [6]. Some researchers tried to use machine learning methods such as AdaBoost and support vector machine (SVM) [7], [8]. Our model is based on spatial-color Gaussian mixture models (SCGMM), which is known to have more discrimination power than color-only models [9]. For prior terms, we have developed pairwise terms imposing the temporal coherence as well as the spatial smoothness. The spatial smoothness terms were based on the contrast of images so that we have smooth object boundaries [6], and the temporal coherence term allows us to have stable segmentation results across frames [6]. An important problem in video segmentation is a model update method. As models for foreground and background are changing as time goes, online learning for models is an essential step for video segmentation. However, it is a chicken-and-egg problem: we need a good segmentation for the update of SCGMM models and the good segmentation requires the accurate SCGMM models. In order to address this problem, we assume that color models are changing slowly while the spatial distribution can change abruptly, and have used the method in [9]. However, we extend the approach so that we can deal with slowly-varying color models. Experimental results have shown that our method is efficient in terms of time and memory and yields stable results compared with conventional methods. The rest of the paper is organized as follows. We present our cost function in Section 2 and explain its automatic initialization and model update methods. Experimental results are provided in Section 3 we conclude this paper in Section 4.

II. Proposed Method

Our method is based on energy minimization framework, and we present our cost function and its minimization method in this section. Although some methods built 3D graphs and solved a 3D energy minimization problem for video segmentation [10], [11], we formulate the video segmentation problem as a frame-by-frame image segmentation problem. It is because we have to satisfy real-time requirements on video-call applications.

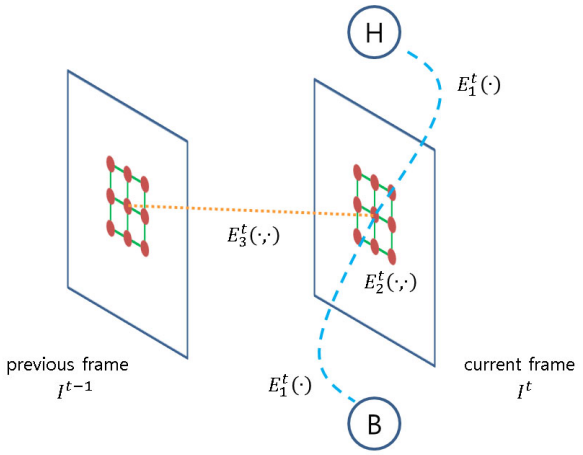

Fig. 1. Illustration of our graph. Red points are pixels and connectivities are represented by lines.

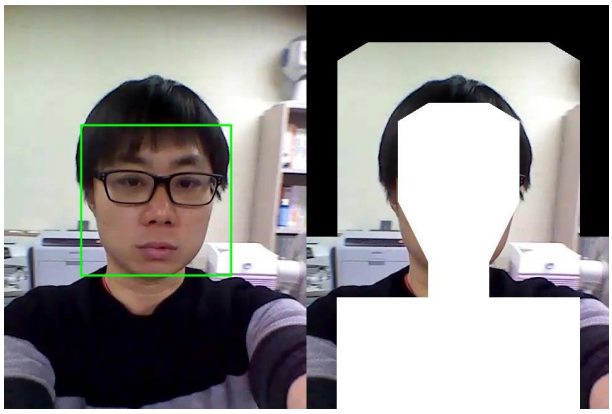

Fig. 2. Region initialization based on face detection results. Left is the face detection result and right shows estimated human/background regions. We represent a human region as white and background as black.

A. Design of our cost function

We denote an image at time t as I^t, and I_i^t means the i-th pixel in the image ($1 \leq i \leq N$). Then our segmentation problem can be considered a problem to assign a label $x_i^t \in \{0,1\}$ to each pixel ($1 \leq i \leq N$) in the t-th frame, where $x_i^t = 0$ means that the i-th pixel is in background region and $x_i^t = 1$ stands for a foreground (human) pixel. We denote our cost function at time t as $E^t(\cdot)$, and we can get the optimal labels by minimizing the cost function:

$$\{\hat{x}_i^t\} = \arg\min_{\{x_i^t\}_{i=1}^N} E^t(\{x_i^t\}). \tag{1}$$

To be precise, the cost function $E^t(\{x_i^t\})$ is given by the sum of three terms

$$\lambda_1 \sum_{i\in\mathcal{V}} E_1^t\left(x_i^t\right) + \lambda_2 \sum_{(i,j)\in\mathcal{E}} E_2^t\left(x_i^t, x_j^t\right) + \lambda_3 \sum_{i\in\mathcal{V}} E_3^t\left(x_i^t, \hat{x}_i^{t-1}\right), \tag{2}$$

where $\mathcal{V}$ is a set of sites, $\mathcal{E}$ is a set of edges in a 4-neighborhood system, and $\hat{x}_i^{t-1}$ is the estimated label in the $(t-1)$-th frame. In our cost function, the first term $E_1^t(\cdot)$ represents the likelihood of being foreground/background pixels, the second term reflects the spatial smoothness constraints, and the last term imposes temporal coherence by preventing abrupt changes in labels across frames.

The data term $E_1^t(\cdot)$ is based on human and background models and we adopt SCGMM for the parametric description of their distributions. We have used 5 dimensional feature vector z_i^t for the i-th pixel, which is the concatenation of 2-dimensional position vector and 3-dimensional color vector (RGB space). We have normalized each element to $[0,1]$. Then, the generative model of a human or background pixel in the t-th frame is given by

$$p_t\left(z_i \mid c\right) = \sum_{k=1}^{M_c} \alpha_{k,c}^t \times G\left(z_i; \mu_{k,c}^t, \Sigma_{k,c}^t\right) \tag{3}$$

where $c \in \{h, b\}$ represents a human or background class. For each class c, $G(\cdot; \mu_{k,c}^t, \Sigma_{k,c}^t)$ is the Gaussian with mean $\mu_{k,c}^t$ and covariance $\Sigma_{k,c}^t$, M_c is the number of Gaussians, and $\alpha_{k,c}^t$ is the weight of the k-th Gaussian. Finally, the data term $E_1^t(\cdot)$ is given by

$$E_1^t\left(x_i^t\right) = \begin{cases} \min\limits_k -\log\left(\alpha_{k,b}^t G(\mathbf{z}_i^t; \mu_{k,b}^t, \Sigma_{k,b}^t)\right) \text{ when } x_i^t = 0 \\ \min\limits_k -\log\left(\alpha_{k,h}^t G(\mathbf{z}_i^t; \mu_{k,h}^t, \Sigma_{k,h}^t)\right) \text{ when } x_i^t = 1. \end{cases} \tag{4}$$

The second term $E_2^t(\cdot,\cdot)$ encodes the smoothness constraints on spatially adjacent labels:

$$E_2^t\left(x_i^t, x_j^t\right) = \frac{1}{d(i,j)} \exp\left(-\frac{\left\|I_i^t - I_j^t\right\|^2}{2\sigma^2}\right)\left|x_i^t - x_j^t\right| \tag{5}$$

where $d(i,j)$ is the distance between two pixels and σ is the empirical standard deviation computed from the pairs in $\mathcal{E}$. Since $E_2^t\left(x_i^t, x_j^t\right)$ becomes large when the adjacent pixels having similar color have different labels, this term helps us to have smooth regions.

Finally, the third term $E_3(\cdot,\cdot)$ imposes temporal coherence in video segmentation. Since the time difference between two consecutive frames are very small in video sequences, many pixels in a current frame should have the same labels to the pixels in the previous frame (especially when they have similar colors). In order to encode this observation, $E_3^t(\cdot,\cdot)$ is given by

$$E_3^t\left(x_i^t, \hat{x}_i^{t-1}\right) = \exp\left(-\frac{\left\|I_i^t - I_i^{t-1}\right\|^2}{2\omega^2}\right)\left|x_i^t - \hat{x}_i^{t-1}\right| \tag{6}$$

where ω is the empirical standard deviation of the color difference between two pixels in the same position but adjacent frames. We have illustrated our graph structure in Fig. 1.

B. Automatic initialization

For the initialization of our method, we should have SCGMM for human and background regions. In usual video-call situations, the frontal face of a user appears in the center of an image, and the Viola-Jones algorithm [12] is applied to the center of video sequences. After face detection, we consider

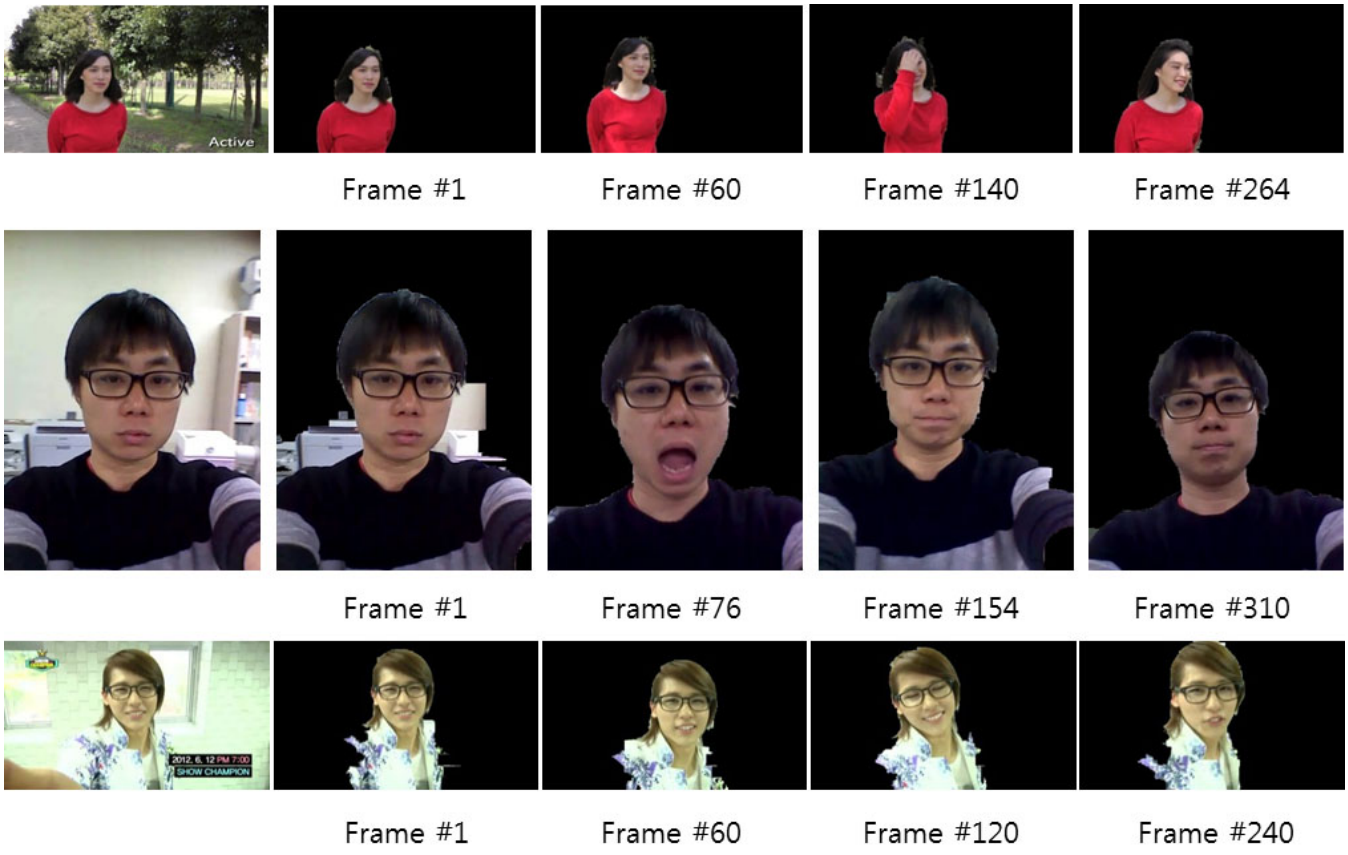

Fig. 3. Human segmentation results.

64% of the face region as seed points and expand them to cover hair and a upper body. When expanding to a upper body, we assume that the width of a body is two times wider than that of a face. Background region is estimated by morphology operations based on estimated human region. Fig. 2 shows our initial labels. From the labels of human and background, we estimate the parameters in SCGMM with the Expectation-Maximization (EM) algorithm. Our algorithm is automatically re-initialized when the human region seems to be poor, e.g., the human region becomes too small.

C. Model updates and segmentation

Since the properties of human and background regions are changing, we have to update models for them. The update of SCGMM parameters is straight-forward when we have additional samples, and one might think that we may use the models in the previous frame (i.e., the labels $\{\hat{x}_i^{t-1}\}$ in the previous frame). However, this idea does not reflect the properties of the current frame well. It is because the spatial distribution can experience dramatic changes between two consecutive frames even though we can assume that the color distributions are slowly changing. In order to alleviate this problem, we have used ideas in [9]: (a) we update the spatial distributions (without the labels in the current frame), (b) estimate the labels in the current frame by minimizing the cost function, and (c) update the color and spatial distributions using the estimated labels.

To be precise, we efficiently update spatial distributions with the EM-algorithm by assuming independence between color and positions. After updating spatial distribution we minimize $E^t(\{x_i^t\}_{i=1}^N)$ with the graph-cut algorithm in order to get the human regions. Since the minimization based on the graph-cut algorithm does not guarantee to yield a connected region [13], we have considered the largest connected component as a foreground region. Then, with the estimated labels, we update SCGMM models. Our method is similar to [9], however, we also update color models in order to deal with slowly-changing colors:

$$\mu_{k,c}^t = (1-\eta)\,\hat{\mu}_{k,c}^t + \eta\ \mu_{k,c}^{current} \tag{7}$$

$$\Sigma_{k,c}^t = (1-\eta)\,\hat{\Sigma}_{k,c}^t + \eta\Sigma_{k,c}^{current} \tag{8}$$

where $c \in \{b, h\}$, η is a constant, $\mu_{i,c}^{current}$ and $\Sigma_{i,c}^{current}$ are a mean and a covariance matrix for the labeled pixels in the current frame, $\hat{\mu}_{k,c}^t$ and $\hat{\Sigma}_{k,c}^t$ are an estimated mean and an estimated covariance matrix that are guessed in the beginning of the current frame after spatial updating. Since it is very time-consuming to use whole pixels in the EM algorithm, we have used the down-sampled image. This approach enables fast update without noticeable degradation.

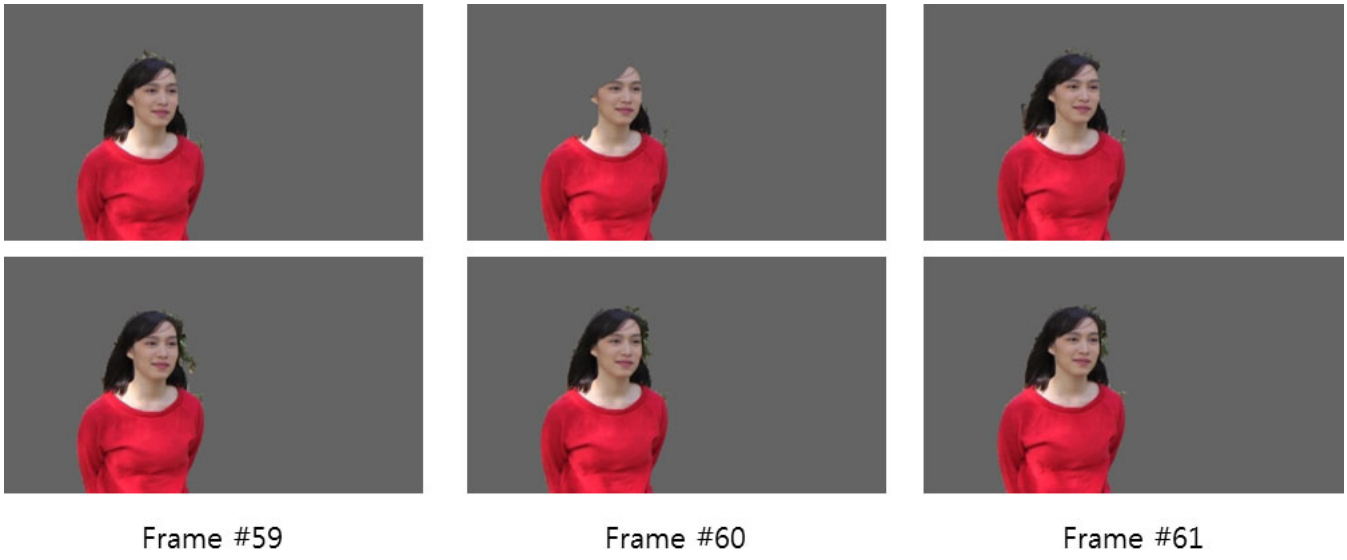

Fig. 4. Comparison to SCGMM tracking [9]. Results of [9] are shown in upper row and our results are in bottom row.

III. Experimental results

We have evaluated our method on several video sequences having approximately 640×480 resolutions. For the automatic initialization of our method, we have used a face detector in [12]. The frame-rate of our method is about 4 (frame/seconds) with a general PC, AMD Phemom(tm) 2 x6 1055T 2.8Ghz. Since our algorithm only requires two adjacent frames for the segmentation, our method is memory-efficient compared with other video-segmentation algorithms [10]. As shown in Fig. 3, our method yields more stable results due to the temporal coherence term and it is efficient due to resized images. Comparison to the SCGMM tracking method [9] is shown in Fig. 4.

IV. Conclusion

In this paper, we have proposed a new human segmentation method for video-call applications. For the automatic initialization, we detect faces in images and initialize spatial-color models for human and background. Based on these models, we extract human regions by minimizing the cost function and the models are updated by the extracted regions. Since our method considers temporal coherence between consecutive frames, our method yields more stable results compared with other methods. Also, our method is memory-efficient and its frame rate is about 4 in VGA processing. Our method is free from restrictions on camera motions, and we believe our method can be employed in other applications.

Acknowledgment

This research was supported by Samsung Electronics, and also by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program NIPA-2013-H0301-13-4005) supervised by the NIPA(National IT Industry Promotion Agency).

References

[1] Yin Li, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum, "Lazy snapping," *ACM Transactions on Graphics (ToG)*, vol. 23, no. 3, pp. 303–308, 2004.

[2] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004.

[3] Yuri Y Boykov and M-P Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *IEEE Conference on International Conference on Computer Vision*, 2001, pp. 105–112.

[4] Yuri Boykov, Olga Veksler, and Ramin Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[5] Antonio Criminisi, Geoffrey Cross, Andrew Blake, and Vladimir Kolmogorov, "Bilayer segmentation of live video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 53–60.

[6] Jian Sun, Weiwei Zhang, Xiaoou Tang, and Heung-Yeung Shum, "Background cut," in *European Conference on Computer Vision*. 2006, pp. 628–641, Springer.

[7] Sang Hak Lee, Hyung Il Koo, and Nam-Ik Cho, "A video object segmentation algorithm based on the feature learning and shape tracking," in *IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 4673–4676.

[8] Minglun Gong and Li Cheng, "Foreground segmentation of live videos using locally competing 1svms," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 2105–2112.

[9] Ting Yu, Cha Zhang, Michael Cohen, Yong Rui, and Ying Wu, "Monocular video foreground/background segmentation by tracking spatial-color gaussian mixture models," in *IEEE Workshop on Motion and Video Computing*, 2007, pp. 5–5.

[10] Yin Li, Jian Sun, and Heung-Yeung Shum, "Video object cut and paste," *ACM Transactions on Graphics (TOG)*, vol. 24, no. 3, pp. 595–600, 2005.

[11] Zhiqiang Tian, Jianru Xue, Nanning Zheng, Xuguang Lan, and Ce Li, "3d spatio-temporal graph cuts for video objects segmentation," in *IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 2393–2396.

[12] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2001, pp. 511–518.

[13] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp, "Image segmentation with a bounding box prior," in *IEEE International Conference on Computer Vision*, 2009, pp. 277–284.