Human Segmentation from Video by Combining Random Walks with Human Shape Prior Adaption

Yu-Tzu Lee*, Te-Feng Su†, Hong-Ren Su*, Shang-Hong Lai†, Tsung-Chan Lee‡ and Ming-Yu Shih‡
*Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu, Taiwan, R.O.C.
E-mail: s9965508@m99.nthu.edu.tw,suhongren@gmail.com
†Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan, R.O.C.
E-mail: tfsu@cs.nthu.edu.tw,lai@cs.nthu.edu.tw
‡Industrial Technology Research Institute, Hsinchu, Taiwan, R.O.C.

***Abstract*—In this paper, we propose an automatic human segmentation algorithm for video conferencing applications. Since humans are the principal subject in these videos, the proposed framework is based on human shape clues to separate humans from complex background and replace or blur the background for immersive communication. We first detect face position and size, track human boundary across frames, and propagate the segmentation likelihood to the next frame for obtaining the trimap to be used as input to the Random Walk algorithm. In addition, we also include gradient magnitude in edge weight to enhance the Random Walk segmentation results. Finally, we demonstrate experimental results on several image sequences to show the effectiveness and robustness of the proposed method.**

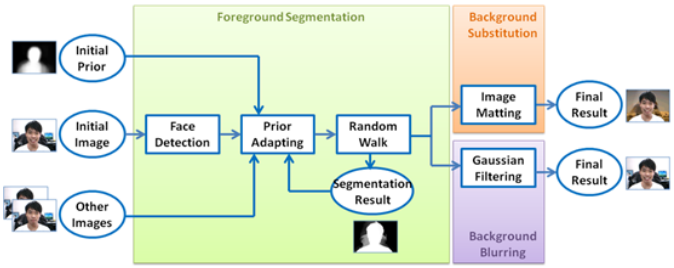


Fig. 1. The flowchart of the proposed system.

I. Introduction

In recent years, video conferencing systems have received more and more attentions in human communication. Using cameras, people can enjoy an immersive communication experience, regardless of distance, that exceeds a telephone call. Therefore, video conferencing systems are extensively used both in corporative domains and for generic purposes, such as communication between friends and family. During video conferencing sessions, different backgrounds are necessary for different purposes. For example, a uniform background is most desirable to protect privacy in a virtual meeting with several participants, and a blurred or suitable shared background may be the preferred option for connecting between friends. To this end, accurately separating the humans from complex backgrounds is a initial, but critical step.

Over the past years, a great amount of effort has been dedicated to this subject. However, attempts to create an efficient and powerful algorithm is still a challenging task. Sun et al. [1] proposed the background cut, which considers both color and contrast as basic model for background subtraction algorithm and resolved the energy function by min-cut algorithm. The background cut produced good results in video sequences with stationary backgrounds and nearly real-time performance. However, it often fail to correctly separate the humans while the background scene is complex or they have similar color. In addition, the learned background assumption always limits the application of the algorithms in real world.

Criminisi et al. [2] proposed a probabilistic fusion formulation that combines motion, color and contrast cues together with spatial and temporal priors to infer binary segmentation. Based on [2], Yin et al. [3], [4] proposed a new motion representation refereed as "motons" that combines both motion and spatial information. They also estimated the likelihood from the spatial context of motion with shape filters in each pixel and efficiently learned the estimation through random forests. Although these approaches achieve good segmentation, they require a large set of manual initialization of the foreground. These limitations are problematic during application to different scene setups.

Several approaches are proposed to improve the foreground segmentation results with human shape template. Zhang et al. [5] used a background subtraction based method as segmentation result and mapped it into a per-pixel blurring radius image to blur the background. Parolin et al. [6] tracked the face to guide a generic Ω-shaped template of the head and shoulders in video sequences. A region of interest (ROI) was created around the generic template and an energy function based on edge, color and motion cues is used to extract the silhouette of humans. Li et al. [7] perform pedestrian segmentation with several trained human shape prior models in Random Walk algorithm. However, a fixed human shape is not flexible enough. Thus, good results are produced only when the user fits the human boundary properly.

Motivated by the above issues, we propose a robust method for upper body segmentation to generate an accurate foreground mask. Firstly, we apply face detection and use a model of the human upper to estimate probability of foreground and background. Next, tracking the human boundary is performed effectively across frames and the segmentation likelihood is

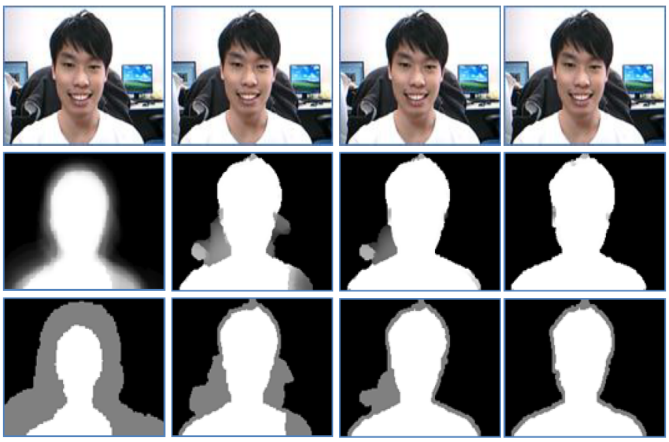


Fig. 2. An example of image sequence for adapting human shape prior and trimap frame by frame. The first row is original sequence, and the second row shows the estimated prior model by prior adaption procedure. The final row shows the seed sets for Random Walk in trimaps (white dots represent foreground seeds and black dots represent background seeds).

propagated to the next frame for obtaining the trimap to be used as input to the Random Walk algorithm. Different from previous methods, using the boundary tracking to update the prior model can reduce the amount of unknowns significantly. The prior probability is also used as a constraint in Random Walk to make the result more temporally coherent. After separating humans from the background, we matte the images by segmentation result to fulfill the background substitution or use Gaussian filter defocusing on the background scenes to synthesize background blurring. Finally, the experimental result is described in terms of segmentation error rate and evaluation efficiency. Fig. 1 illustrates the flowchart of the proposed method.

The rest of this paper is organized as follows. In Section II, we introduce the proposed adaptation of human shape prior model, and separating humans from the background with the shape prior in Random Walk algorithm is described in Section III. In Section IV, we describe two applications after separating humans from background. In Section V, we show some experimental results in terms of segmentation error rate and evaluation efficiency. Section VI concludes this paper.

II. Human Shape Prior Adaption for Video

In the framework of the proposed method, human shape prior plays an important role for foreground segmentation. Better human shape prior model can reduce mistakes from cluttered background, and we can focus on features extracted only along human boundary in human object segmentation. Unfortunately, in order to apply the shape prior to different human shapes in each video frame, the boundary of the human shape prior model usually coarsely close to ideal human shape. The large number of unknown pixels between foreground and background seeds, which comes from coarse shape human prior, cause the increase of computational cost in Random Walk algorithm and inaccurate silhouette extraction of humans.

Fig. 3. An Example for generating boundary pixels for new shape. (a) The segmentation result from previous frame. (b) Green dots represent foreground pixels and blue dots represent background pixels (c) Correspond relation between continuous two frames defined by tracker. (d) Propagate corresponding likelihood from previous frame to the current human shape prior.

Based on the above concept, we implement human shape prior adaption procedure to make the boundary of human shape prior model closer to the input image both in the probability map and the trimap frame by frame. Fig. 2 depicts an example of human shape prior adaption. It can improve the segment performance and reduce execution time simultaneously in the Random Walk algorithm.

A. Boundary Tracking

To generate a more fit shape prior for the current human object in video, we need to estimate human motion from video. Optical flow is commonly used to estimate the motion of every pixel, but its complexity is very high. We only need the motion at the pixels that are related to the human shape, thus we choose to find the motion by tracking the boundary pixels. The boundary pixels are defined to be the pixels with associated probability values between the lower and upper bounds, which correspond to the thresholds for deciding the background and foreground pixels from the probability map. Fig. 3 illustrates an example of generating boundary pixels for new shape. Specifically, we compute the motion vector (d_u, d_v) at boundary pixel (u, v) between two adjacent frames. We use the Lucas-Kanade tracker [8] to estimate the motion at the boundary pixels in consideration of the accuracy and efficiency.

B. Gray Area Determination

To determine the gray area between the foreground (human) and background regions as the unknown pixels in the Random Walk segmentation, we apply two techniques: namely the gradient based pixel growing and region expansion. The gradient based method use the similarity in the gradient magnitude and direction in the neighborhood of boundary pixels to grow the boundary region based on the following criteria:

$$|\nabla e(u,v) - \nabla e(u_{neibor}, v_{neibor})| \leq E \tag{1}$$

$$|m(u,v) - m(u_{neibor}, v_{neibor})| \leq M \tag{2}$$

where $\nabla e(u,v)$ and $m(u,v)$ are the image gradient and its magnitude at pixel (u,v). In addition, the gradient based pixel growing and region expansion methods are directly used to

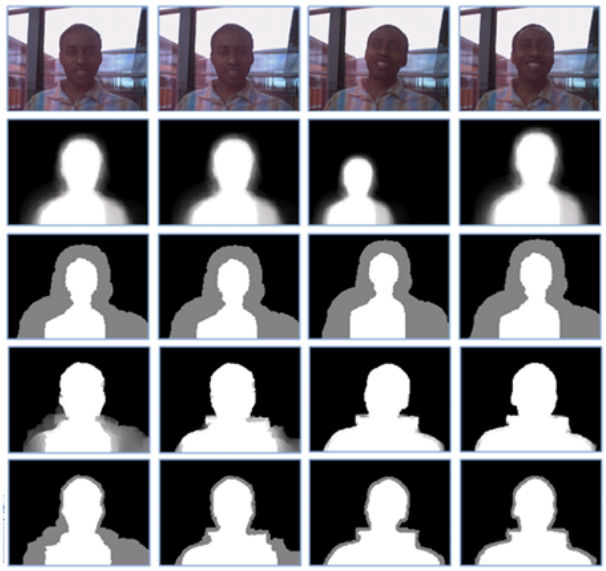


Fig. 4. Intermediate results for prior adaptation. The first row is original sequence. The second and third rows come from baseline method, while the last two rows come from prior adaption. In samples based on different methods above, the upper row shows the estimated priors, and the lower row shows the trimaps which are the fg/bg seeds in Random Walk.

expand the boundary region with a fixed width, which is accomplished by the dilation operator.

C. Trimap Construction for Random Walk Segmentation

After the above steps, the shape prior model is adapted to the input frame. We can simply apply connected component analysis to label the trimap, which include the definite background, definite foreground and the boundary region. In Fig. 4, we compare the human shape prior models and the trimaps obtained by using the baseline method (only aligned initial prior model by face detection) and our prior adaption procedure. The white pixels and black pixels are the initial seeds for the Random Walk segmentation algorithm, while the gray pixels denote the unknown pixels to compute in the segmentation algorithm.

III. Foreground Segmentation

In this section, we describe the human segmentation in modified Random Walk algorithm with human shape prior model. Random Walk algorithm is a graph-based image segmentation method which assign each pixel a label of foreground or background when for an image is given. A graph $G = (V, E)$ has vertices $v \in V$, with each vertex corresponding to a pixel and $V = \{v_i\}_{i=1...N}$. A weighted graph has a value assigned to each edge, and it is called a weight. The weight between two vertices v_i and v_j is denoted by w_{ij}.

Similar to [7], the human shape prior model are combined into a single energy function with the introduction of a free parameter that controls the weighting between the two energy functions as follows.

$$E(\Omega) = E_{RandomWalks} + \lambda E_{Prior} \tag{3}$$

whereλ is a weighting coefficient. The $E(\Omega)$ is the sum of two terms : $E_{RandomWalks}$ and E_{Prior}. The first term $E_{RandomWalks}$ is the label-continuity constraint that two neighboring pixels in the small neighborhood system should have the same label if their colors or intensities are similar.

$$E_{RandomWalks} = \sum_{e_{ij} \in E} w_{ij}^{total} (x_i - x_j)^2 \tag{4}$$

Different from original weighting, we combine the edge weights determined from the intensity and gradient information into the total edge weight w_{ij}^{total}.

The second term E_{Prior} is the constraint that each pixel tends to the shape prior model and formulated as follows.

$$E_{Prior} = \sum_{v_{ij} \in V} D_i (x_i - p_i)^2 \tag{5}$$

where D_i is the importance of the node which is determined by the multiplying the associated degree of the prior model and the degree of the same node determined by its weight. x_i and p_i represent the likelihood and probability at node v_i of the shape prior model, respectively. The energy function can be minimized in Random Walk algorithm [9].

IV. Applications related to Human Segmentation

A. Background substitution

In order to map the likelihood $\overrightarrow{x}$ to a background substitution image we use the image matting formula given by

$$I[u, v] = \alpha_i * I[u, v] + (1 - \alpha_i) * I[u, v] \tag{6}$$

where α_i corresponds to the probably x_i for a pixel i to be a foreground label computed from the Random Walk algorithm described in the previous subsection.

B. Background blurring

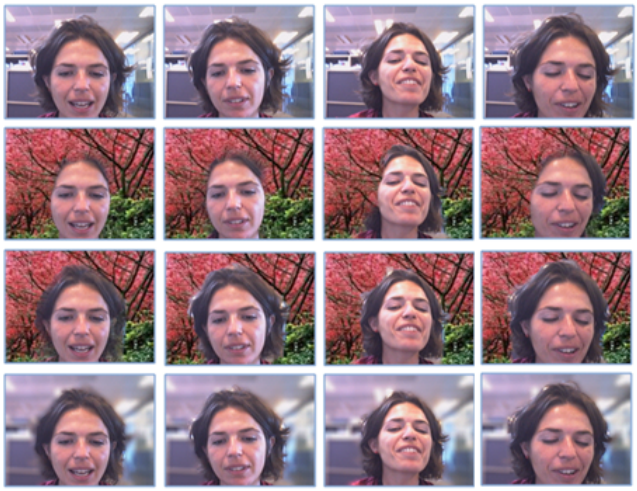
Background blurring is useful to preserve privacy efficiently in the video conferencing. In order to apply the likelihood $\overrightarrow{x}$ to select the background region for image blurring, we apply a Gaussian filter in the background region as follows:

$$I[u, v] = \left\{ \begin{array}{c} I[u, v] * G_\sigma(u, v), \text{if} x_i < \tau \\ I[u, v], \text{otherwise} \end{array} \right\} \tag{7}$$

where the image $I[u, v]$ is convolved with the Gaussian kernel G_σ. We use $wz = 15$ and $\sigma = 7$ and obtain good results in both indoor and outdoor background blurring in this paper.

V. Experimental results

In this section, we show the results of background substitution/ blurring by using the proposed method and give performance comparison with previous methods. Our system was implemented in C++ program with openCV library on a PC which is equipped with Intel Core2 CPU 6320 running at 1.86 GHz processor and with 4GB RAM. The the probability map of shape the prior model was estimated by taking averages for all the silhouettes which were extracted from 284 images from 14 human videos by using background subtraction.



(a) Some frames of the ”41” sequence.

(b) Some frames of the ”54” sequence.

Fig. 5. Experimental results by using the proposed algorithm and the baseline method on the videos from dataset [12]. In each sequence, the first row shows the original image, the second row give the background substitution results by the baseline method. The third and fourth rows present the background substitution and background blurring results from the proposed approach, respectively.

The experiment is performed by applying the proposed algorithm on several videos from dataset [3], with each video containing 49∼500 testing frames of size 240x320. After separating human from the background, we matte the images by segmentation result to fulfill the background substitution or use Gaussian filter defocusing on the background scenes to synthesize background blurring. Fig. 5 show the results of the proposed prior adaption scheme and the baseline method which only used face detection in conjunction with a trained human shape prior. We also evaluate our work both on performance and execution time and demonstrate the effectiveness of our approach in Table I.

As evident from the experimental results, the proposed approach can reduce the segmentation error rate and the execution time at the same time. The different of execution time among videos is mainly due to different sizes of targets or the number of unknown pixels in the random walks segmentation. For example, when the target changes the pose too quickly, it makes large changes in intensity, which will make the LK tracker fail in adapting the human shape prior model.

TABLE I
THE SEGMENTATION ERROR FOR DIFFERENT SEQUENCES BY USING THE BASELINE METHOD AND THE PROPOSED METHOD

Sequence	frames	baseline		our method	
		Error	fps	Error	fps
21	66	2.43	2.097	1.76	12.002
41	317	6.23	1.919	0.82	9.622
51	364	2.39	2.046	1.26	14.016
54	530	8.40	1.579	1.70	10.793
58	330	16.26	2.228	0.56	14.315
Average	321.4	7.142	1.974	1.22	12.150

VI. CONCLUSIONS

In this paper, we presented an automatic human segmentation algorithm from video for applications based on both spatial and temporal cues. Our experimental results indicate that the proposed approach produces human segmentation with temporal and geometric coherence because of the adapted human shape prior model. In the future, we intend to collect more human upper bodies with different poses and construct a more complete human shape model. Furthermore, using the GPU computing may help to significantly reduce the execution time, so it is a worthwhile direction for developing a real-time video conferencing system.

ACKNOWLEDGMENT

This work was supported in part by ITRI Advanced Research Program from the project 102-EC-17-A-01-05-0337.

REFERENCES

[1] J. Sun, W. Zhang, X. Tang, and H.-Y. Shum, “Background cut,” in *Proceeding of European Conference on Computer Vision (ECCV)*, 2006.
[2] A. Criminisi, G. Cross, A. Blake, and V. Kolmogorov, “Bilayer segmentation of live video,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
[3] P. Yin, A. Criminisi, J. Winn, and I. Essa, “Tree-based classifiers for bilayer video segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
[4] P. Yin, A. Criminisi, J. Winn, and I. Essa, “Bilayer segmentation of webcam videos using tree-based classifiers,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 30 –42, jan. 2011.
[5] C. Zhang, Y. Rui, and L.-W. He, “Light weight background blurring for video conferencing applications,” in *IEEE International Conference on Image Processing (ICIP)*, 2006.
[6] A. Parolin, G. P. Fickel, C. R. Jung, T. Malzbender, and R. Samadani, “Bilayer video segmentation for videoconferencing applications,” in *IEEE International Conference on Multimedia and Expo (ICME)*, 2011.
[7] H.-R. Su K.-C. Li and S.-H. Lai, “Pedestrian image segmentation via shape-prior constrained random walks,” in *Proceeding of Pacific-Rim Symposium on Image and Video Technology (PSIVT)*, 2011.
[8] B. D. Lucas and Takeo Kanade, “An iterative image registration technique with an application to stereo vision,” in *Proceedings of Imaging Understanding Workshop*, 1981.
[9] L. Grady, “Random walks for image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1768 –1783, nov. 2006.