Automatic Human Body Segmentation using Level-Set based Active Contours followed by Optical Flow in Video Surveillance

Muhammad. Hameed Siddiqi
Ubiquitous Computing Lab, Kyung Hee University, Korea
siddiqi@oslab.khu.ac.kr

Phan Tran Ho Truc
Ubiquitous Computing Lab, Kyung Hee University, Korea
pthtruc@oslab.khu.ac.kr

Sungyoung Lee
Ubiquitous Computing Lab, Kyung Hee University, Korea
sylee@oslab.khu.ac.kr

Young-Koo Lee
Ubiquitous Computing Lab, Kyung Hee University, Korea
yklee@khu.ac.k

***Abstract*—Human body segmentation is a critical module in video-based activity recognition (AR) because it defines the image area necessary and sufficient for the follow-up modules like feature extraction. Existing methods often involve modeling of the human body and/or the background, which normally requires extensive amount of training data and cannot efficiently handle changes over time. Recently, active contours have been emerging as an effective segmentation technique in still images. In this paper, an active contour model is adapted that is robust to illumination and clothing changes, typical issues in practical AR systems. To make the model work smoothly with video data, the optical flow is used, which is estimated in two consecutive frames, to position the initial contour in the current frame. The proposed approach is unsupervised, i.e., no training data or prior human model is needed. The proposed model gives prominent results of segmentation.**

Keywords; video surveillance; body segmentation; active contour; optical flow

I. INTRODUCTION

Recognizing human activities is an increasingly active research area and is a key component in many computer vision and pattern recognition applications, such as video surveillance. The accuracy of the video-based human activity recognition depends significantly on the performance of human body segmentation. Existing video segmentation methods can be roughly categorized using three types of image measurements on which the segmentation is based, i.e., motion, appearance, or shape. The principle of motion-based segmentation is that by detecting the motion from consecutive images, one can find the human. The motion can be measured using either flow estimation [1-3] or image differencing [4, 5]. But in these methods, if the body parts that have not moved nor have image intensity similar to their neighbors' will not be detected that is one of the limitations of this category. Appearance-based segmentation is based on a simple assumption that appearance of human is different from that of the background. Methods in this category work by first constructing a human appearance model and then extracting pixels in the current image that match with the constructed model. Those methods not only extract human from background but can also distinguish individuals from one another by building up a distinct appearance model for each individual [6-8]. Similar to appearance-based approaches, the shape-based segmentation is built on the idea that the shape of a human is different from that of other objects in an image. However, the shapes of individuals are almost the same, making the approaches in this category suitable for tracking simple correspondences [9-12]. Appearance-based and shape-based segmentation methods require supervised learning, i.e., a model of each human has to be built up in advance to train a suitable classifier or to compare with the model of the segmented foreground objects. Building up models able to handle changes over time remains an open issue. On one hand, a model should adapt quickly to change, but on the other hand, long term temporal consistency is required.

In this research, a model is proposed that use unsupervised segmentation in combination with motion information. Specifically, in each video frame, an active contour (AC) [13] is evolved to capture the human body and the motion information (i.e., optical flow) is used to move the contour toward the human position in the next frame, where the contour is again evolved to detect the exact boundary of the body that avoid the need of supervised learning. In AC models, the initial contour should be close to the object in order to converge correctly. This can be done manually in static images but will not be feasible in video data which have a large number of frames. Automating the initialization of AC is therefore needed. Optical flow is a good candidate for this purpose because it contains the relatively exact direction and magnitude of the motion between consecutive frames. According to the afore-mentioned classification, the proposed approach falls into the shape-based category with partial incorporation of temporal context. By "partial", means that the temporal information is not embedded in the contour evolution itself but instead is used as a guideline for positioning the initial contour.

We already discussed some related work about the human body segmentation. The rest of the paper is organized as follows. The nest two sections provide an overview of the proposed approach, and the experimental results and discussion on its comparison with the conventional Chan-Vese AC in two conditions: without and with optical flow incorporation. In the last section, the paper will be concluded after some discussions.

This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency)" (NIPA-2010-(C1090-1021-0003)).

II. METHODOLOGY

The accuracy of the video-based human activity recognition depends significantly on the performance of human body segmentation. In the field of image segmentation, since it was first introduced by [13], active contour (AC) model has attracted much attention. Recently Chan and Vese (CV) proposed in [14] a novel form of AC that does not use the edge information (as the other AC models use), but utilizes the differences between the regions inside and outside of the curve. Its energy functional is defined as:

$$F(C) = \int_{in(C)} |I(x) - c_{in}|^2 \, dx + \int_{out(C)} |I(x) - c_{out}|^2 \, dx \quad (1)$$

where $x \in \Omega$ (the image plane) $\subset R^2, I : \Omega \to Z$ is a certain image feature such as intensity, color, or texture, and c_{in} and c_{out} are respectively the mean values of image feature inside [*in(C)*] and outside [*out(C)*] the curve *C*, which represents the boundary between the two separate segments. The global minimization of the above energy functional does not provide better result when segment is inhomogenous as shown in Figure 1(b). The proposed methodology contains, to incorporate an evolving term based on the Bhattacharyya distance to the CV energy functional such that not only the differences within each region are minimized but the distance between the two regions is maximized as well. The proposed energy functional is:

$$E_0(C) = \beta F(C) + (1-\beta) B(C)$$

where $\beta \in [0,1]$, $B(C) \equiv B = \int_Z \sqrt{p_{in}(z)\, p_{out}(z)}\, dz$ the Bhattacharyya coefficient [15] with $\sqrt{}$

$$p_{in}(z) = \frac{\int_\Omega \delta(z - I(x)) H(-\phi(x))\, dx}{\int_\Omega H(-\phi(x))\, dx}$$

$$p_{out}(z) = \frac{\int_\Omega \delta(z - I(x)) H(\phi(x))\, dx}{\int_\Omega H(\phi(x))\, dx}$$

$\phi: \Omega \to R$ the level set function, and $H(\bullet)$ and $\delta(\bullet) \triangleq H'(\bullet)$ respectively the Heaviside and the Dirac functions [14]. Note that the Bhattacharyya distance is defined by [*-log B(C)*] and the maximization of this distance is equivalent to the minimization of *B(C)*. Note also that to be comparable to the *F(C)* term, *B(C)* is multiplied by the area of the image because its value is always within the interval [0,1] whereas *F(C)* is calculated based on the integral over the image plane. In general, we can regularize the solution by constraining the length of the curve and the area of the region inside it. Therefore, the energy functional is defined by:

$$E(C) = \gamma \int_\Omega |\nabla H(\phi(x))|\, dx + \eta \int_\Omega H(-\phi(x))\, dx + \beta F(C) + (1-\beta) B(C) \quad (2)$$

where $\gamma \geq 0$ and $\eta \geq 0$ are constants.

The intuition behind the proposed energy functional is that we seek for a curve which 1) is regular (the first two terms) and 2) partitions the image into two regions such that the differences within each region are minimized (i.e., the *F(C)* term) and the distance between the two regions is maximized (i.e., the *B(C)* term). The level set implementation for the energy functional in (2) can be derived as:

$$\frac{\partial \phi}{\partial t} = |\nabla \phi| \left\{ \begin{array}{l} \gamma k + \eta + \beta \left[(I - C_{in})^2 - (I - C_{out})^2 \right] - \\ (1-\beta) \left[\begin{array}{l} \frac{B}{2} \left(\frac{1}{A_{in}} - \frac{1}{A_{out}} \right) + \\ \frac{1}{2} \int_\Omega \delta(z - I) \left(\frac{1}{A_{out}} \sqrt{\frac{p_{in}}{p_{out}}} - \frac{1}{A_{in}} \sqrt{\frac{p_{out}}{p_{in}}} \right) dz \end{array} \right] \end{array} \right\}$$

where A_{in} and A_{out} are respectively the areas inside and outside the curve *C*.

As a result, the proposed model can overcome the CV AC's limitation in segmenting inhomogeneous objects as shown in Fig. 1(c), yielding the body detector more robust to illumination changes and clothing.

(a) (b) (c)

Figure 1: Sample segmentation of inhomogeneous body-shape object using active contours. (a) Initial contour, (b) segmentation result of CV AC, and (c) proposed approach. The CV AC fails to capture the whole body where as the proposed approach succeeds

A. Optical Flow

The convergence of active contour models generally depends on the initial contour [14] that is usually done manually in static image segmentation. In this paper, a model is proposed that incorporates motion information associated in video surveillance to automat this process that can be effectively obtained using optical flow, which is defined as the pattern of motion of objects in a visual scene.

In this paper, a recent implementation of optical flow presented in [16] is used, which was based on [17, 18]. The main idea was to integrate the warping technique [19, 20] into a variational framework where the proposed energy functional consists of non-linearized constraints on intensity constancy, gradient constancy, and smoothness. The coarse-to-fine warping technique helps to implement the optimization of non-linearized constancy assumptions and provide a better optical flow estimation in case of large displacements, i.e., more than one pixel per frame. The accuracy of this model is therefore significantly higher than that of other methods in the literature.

III. RESULTS AND DISCUSSIONS

This paper incorporates motion information associated in video surveillance that can be effectively obtained using optical flow. In order to evaluate the proposed algorithm, a publicly available dataset [21] is used, which consists of ten activities, and each activity was performed by nine different people. The frame size is 144 x 180.

A. Segmentation without Optical Flow

In video surveillance, the active contour evolution in a certain frame is performed independently of the other frames,

means that the human body segmentation in video is done frame-based. The only utilized information is the final contour obtained in the previous frame which will be used to determine the initial position of the active contour in the current frame. First, an ellipse with major axis along *y*-axis of length 25 and minor axis along *x*-axis of length 10 is selected as the initial contour. In the experimental results of this paper, this initial shape will be same for all frames, but only center location varies. In each video, the first frame is segmented using manual initialization such that the initial contour is closer to the object.

Then from the second frame, the position of the initial contour's center in the current frame is the mean value of the points along the final contour in the previous frame. For example, suppose that along the final contour of frame $n (n \geq 1)$, there are M points $\left(x_i^{(n)}, y_i^{(n)}\right)$, $i = 1..M$. Then, the center $\left(c_x^{(n+1)}, c_y^{(n+1)}\right)$ of the initial contour in frame $(n+1)$ is calculated as:

$$c_x^{(n+1)} = \frac{1}{M}\sum_{i=1}^{M} x_i^{(n)} \; ; \quad c_y^{(n+1)} = \frac{1}{M}\sum_{i=1}^{M} y_i^{(n)}$$

Some experimental results of image segmentation of different video activities are shown in Fig. 2, which indicates that the proposed model with the above-described scheme works well with static activities like bend, wave, or jack, but for dynamic activities such as run, walk, or skip, it fails to capture the whole body correctly.

B. Segmentation with Optical Flow

The proposed algorithm fails to segment the human body in dynamic activities, because the reason is that there is a large displacement of object between consecutive frames of those videos, making the previous-frame-based initial position in the current frame far from the object of interest. To overcome this problem, we propose to incorporate the optical flow to move the initial contour closer to object, i.e., in each video, the first frame is segmented using manual initialization exactly the same as in the previous experiment.

From the second frame, $\left(x_i^{(k)}, y_i^{(k)}\right)$, $i = 1..N$, be N points along the final contour of the frame $k, (k \geq 1)$ $\left(v_x^{(k+1)}(x, y), v_y^{(k+1)}(x, y)\right)$ be the optical flow of the frame $(k+1)$ at the point $(x, y) \in \Omega$, where Ω is the frame plane.

Then, the center $\left(c_x^{(k+1)}, c_y^{(k+1)}\right)$ of the initial contour in frame $(k+1)$ calculated as:

$$c_x^{(k+1)} = \frac{1}{N}\sum_{i=1}^{N} x_i^{(k+1)} \; ; \quad c_y^{(k+1)} = \frac{1}{N}\sum_{i=1}^{N} y_i^{(k+1)}$$

where $x_i^{(k+1)} = x_i^{(k)} + v_x\left(x_i^{(k)}, y_i^{(k)}\right); y_i^{(k+1)} = y_i^{(k)} + v_y\left(x_i^{(k)}, y_i^{(k)}\right)$

The idea is that all points along the final contour in the previous frame will be shifted accordingly to the optical flow of the current frame and the mass center of those new points will be used as the center of the initial contour in the current frame. Fig. 3 shows a sample of this initialization scheme.

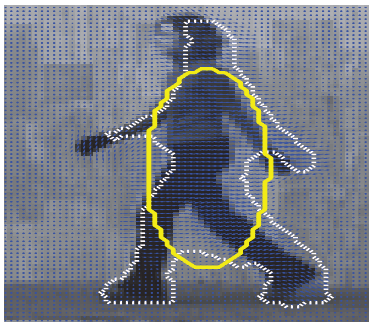

Fig. 3. Initialization using optical flow. White dot line: final contour in the previous frame; blue arrows: optical flow; yellow solid ellipse: initial contour for the current frame.

The results of the proposed model with dynamic activities are compared with CV AC model with and without optical flow are shown in Fig. 4, 5 and 6 respectively.

It is to be noted from Fig. 6, that the segmentation results of the proposed model with the described initialization. It is possible to see that the images from both "static" and "dynamic" activity videos were correctly segmented.

IV. CONCLUSION

This paper has presented an active contour model for human body segmentation from video data that is the modefied model of [22]. Compared to the conventional CV AC for static image segmentation, the proposed model is more robust to noise, illumination changes, and clothing, when applied to video data. The prposed AC model incorporates the motion information which can be effectively estimated using optical flow technique. The optical flow is used to shift the final contour in the previous frame toward the object in the current frame. The mass center of the shifted contour is then used as the center of the initial contour in the current frame. Because optical flow has significant values on parts that move much and is almost zero on static parts of the body, contour points will move or stand still accordingly. As a result, the proposed AC model with this initialization scheme can correctly segment human body in both "static" and "dynamic" activity videos, The comparison results with conventional CV AC are shown in Fig. 4, 5 and 6 respectively

REFERENCES

[1] H. Sidenbladh, "Detecting human motion with support vector machines," in ICPR, Cambridge, 2004

[2] P. Sangi, J. Heikkila, and O. Silven, "Extracting motion components from image sequences using particle filters," in The 12th Scandinavian Conference on Image Analysis, Bergen, Norway, 2001

[3] G.R. Bradski and J.W. Davis, "Motion segmentation and pose recognition with motion history gradients," Machine Vision and Applications, vol. 13, no. 3, pp. 174-84, 2002

[4] I. Haritaogl, D. Harwood, and L.S. Davis, "W4: real-time surveillance of people and their activities," IEEE Trans. PAMI, vol. 22, no. 8, pp. 809-830, 2000

[5] P. Viola, M.J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," IJCV, vol. 63, no. 2, pp. 153-61, 2005

[6] I.B. Ozer and W.H. Wolf, "A hierarchical human detection system in (Un) compressed domains," IEEE Transactions on Multimedia, vol. 4, no. 2, pp. 283-300, 2002

[7] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," IEEE Trans. PAMI, vol. 25, no. 5, pp. 564-75, 2003

[8] C. Stauffer and W.E.L. Grimson, "Learning patterns of activity using real-time tracking," IEEE Trans. PAMI, vol. 22, no. 8, pp. 747-57, 2000

[9] L. Zhao and C.E. Thorpe, "Stereo- and neural network-based pedestrian detection," IEEE Trans. PAMI, vol. 1, no. 3, pp. 148-54, 2000

[10] Y. Wu and T. Yu, "A field model for human detection and tracking," IEEE Trans. PAMI, vol. 28, no. 5, pp. 753-65, 2006

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in CVPR, San Diego, 2005

[12] shape models for tracking non-rigid objects," Pattern Recognition Letters, vol. 24, pp. 1751–65, 2003

[13] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models," Int. J. Comput. Vis., vol. 1, pp. 321-31, 1988

[14] T. Chan and L. Vese, "Active contours without edges," IEEE Trans. Image Proc., vol. 10, pp. 266-77, 2001

[15] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," IEEE Trans. Commun. Technol. , vol. 15 , pp. 52–60, 1967

[16] C. Liu, "Beyond Pixels: Exploring New Representations and Applications for Motion Analysis," Massachusetts Institute of Technology, Ph.D. Thesis 2009

[17] T. Brox, A. Bruhn, N. Papenberg, and J.Weickert, "High accuracy optical flow estimation based on a theory for warping," in European Conference on Computer Vision, 2004, pp. 25-36

[18] A. Bruhn, J. Weickert, and C. Schnorr, "Lucas/Kanade meets Horn/Schunk: combining local and global optical flow methods," Int. J. Comp. Vis., vol. 61, no. 3, pp. 211-31, 2005

[19] M. J. Black and P. Anandan, "The robust estimation of multiple motions: parametric and piecewise smooth flow fields," Comp. Vis. Image Und., vol. 61, no. 3, pp. 75-104, 1996

[20] E. Memin and P. Perez, "A multigrid approach for hierarchical motion estimation," in Proc. Sixth Int. Conf. Comp. Vis., 1998, pp. 933-8

[21] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes," IEEE Trans. PAMI, vol. 29, no. 12, pp. 2247-53, 2007

[22] M. H. Siddiqi, P. T. H. Truc, S. Y. Lee, and Y.–K. Lee, "Level Set Based Automatic Human Body Segmentation," in Proc. 11th International Conference on Pattern Recognition and Information (PRIP'11), Minsk, Belaruse, 2011.

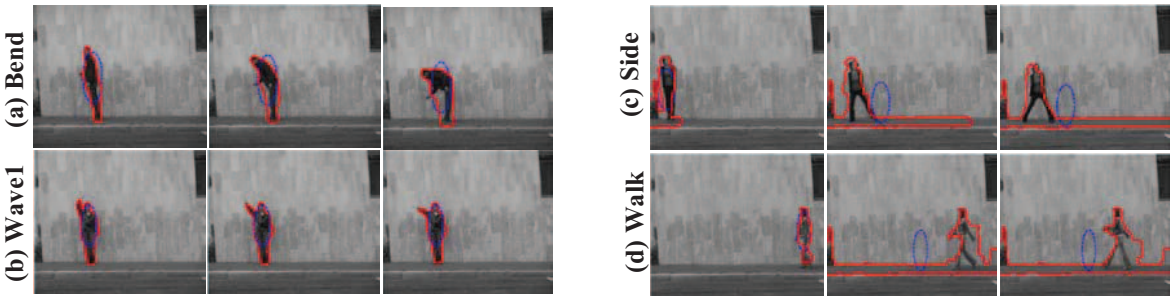

Fig.2. Sample segmentation results of the proposed model without optical flow. Blue dot eclipse: initial contour and red solid curve: final contour representing the segmented object. The model works with static activities like "bend" or "wave", but fails to capture the correct human body in dynamic activities like "side" or "walk"

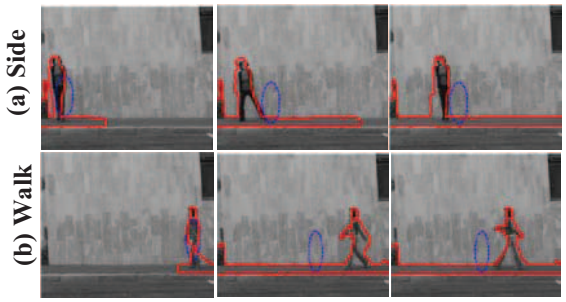

Fig.4. Sample segmentation results of CV AC without optical flow. Blue dot eclipse: initial contour and red solid curve: final contour. The whole body cannot be detected correctly

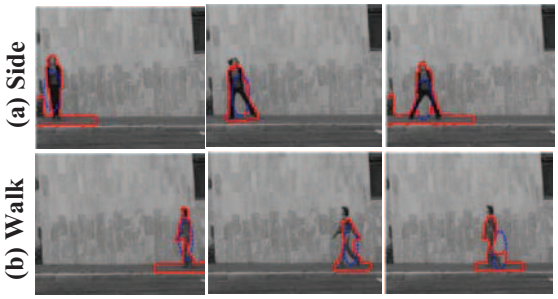

Fig.5. Segmentation results of CV AC model with optical flow. The CV AC fails to capture the correct body even with optical flow incorporation