Upper Body Human Detection and Segmentation in Low Contrast Video

Ruofeng Tong, Di Xie, and Min Tang

***Abstract*—In the application of extracting human regions from videos, many existing methods may lose their efficacy when illumination varies or the human remains still. To address this problem, we propose a method in this paper for human region detection and segmentation by constructing a generalized human upper body model. The method mainly consists of two main procedures. Firstly, foreground connected regions are extracted by background subtraction from current frame and classified through a human upper body model pre-trained with SVM (Support Vector Machine) to determine whether they are human regions. Secondly, we assign an energy function to the region contour and apply an energy minimization procedure to evolve the contour when human regions are "polluted" by background, for example, changes of lighting conditions. After finding the optimal contour we update the background and repeat the procedures in next frame. This feedback strategy rectifies the mistaken background regions promptly and extracts human regions correctly. Our experimental results demonstrate that the proposed method is robust enough to handle videos of low contrast as well as normal conditions.**

***Index Terms*—Human detection, Segmentation, Shape feature, SVM classification, Energy minimization.**

I. Introduction

AUTOMATIC human detection and segmentation in videos are two key procedures of various surveillance applications. Methods of human detection usually find the foreground object from videos and identify them as human or non-human based on shape [1], [2], [3], color [4], [5] or other features [6]. Background subtraction is a common preprocess technique to extract foreground region. Other methods which are based on machine learning introduce classifiers generated by new exploited features. Gradient-based features [7] are the most representatives among them. These methods get rid of background subtraction but have an expensive computation cost which hinders their application in realtime systems. Video segmentation methods are also based on background subtraction techniques while integrated with probabilistic framework, such as Bayesian theory and Markov chain Monte Carlo.

Manuscript received April 10, 2012; revised August 25, 2012. This work was supported by the National Basic Research Program (No. 2011CB302205) of China, the National High-Tech Research and Development Program (No. 2013AA013903) of China, the Science and Technology Program (2010C13023) of Zhejiang Province, China and Zhejiang Provincial Natural Science Foundation of China (No.Y1100018).

Ruofeng Tong is with CAD&CG state key Lab, AI Institute, Zhejiang University, Hangzhou, China (e-mail: trf@zju.edu.cn).

Di Xie is with Department of Computer Science, Zhejiang University, Hangzhou, China.

Min Tang is with CAD&CG state key Lab, AI Institute, Zhejiang University, Hangzhou, China.

Many methods may act dysfunctional since they rely on a good background subtraction result, which can be invalid in case of illumination variance. Some methods may be adaptive to the changes, however, they fail to work well in cases when foreground objects keep still for long and gradually become background. Moreover, many surveillance systems practically equipped with low quality cameras yield low contrast videos, which makes the problem even tougher.

To attempt the challenge, we propose a novel method combining the subtraction-based one with SVM machine learning algorithm. A new feature set, similar to shape contexts [8] but simpler, is also introduced. It is evaluated by a polar coordination and transformed to a 2D histogram. Besides the new shape feature, we also propose a second procedure which utilizes energy function minimization to overcome the impact of illumination variation. The wrong contour caused by illumination variation will be discarded and the contour in the previous frame will evolve when the minimization procedure proceeding and at last converge to a right one like the active contour algorithm [9]. Then we use the evolved contour to update background template. The two procedures are mutually improved with each other and accomplish a good performance collaboratively since the first provides an ideal initial contour and the second back feeds a reliable foreground region.

The rest of this paper is organized as follows. Sections II and III introduce related work and the overview of our algorithm respectively. Section IV demonstrates the SVM based training and detection method as well as the feature set we use. Section V describes the energy function minimization method. Section VI discusses experimental results. Section VII, the final, concludes the paper.

II. Related Work

A. Human Detection

There are many researches focusing on human whole body detection. Recent researches focus on detection with a variety of learning algorithms and features. Volia et al. [10] applied Haar-like features from both appearance and motion in AdaBoost learning but the method mainly detects human at very small scales. Dalal and Triggs [7] used HOG (Histogram of Oriented Gradients) features in training support vector machine while Hou et al. [11] utilized EHOG features with a detector using Vector Boosting. The EHOG features are formed via dominant orientations in which gradient orientations are quantified into several angle scales that divide gradient orientation space into a number of dominant orientations. Blocks of combined rectangles with their dominant orientations constitute the feature pool. Maji et al. [12] improved the

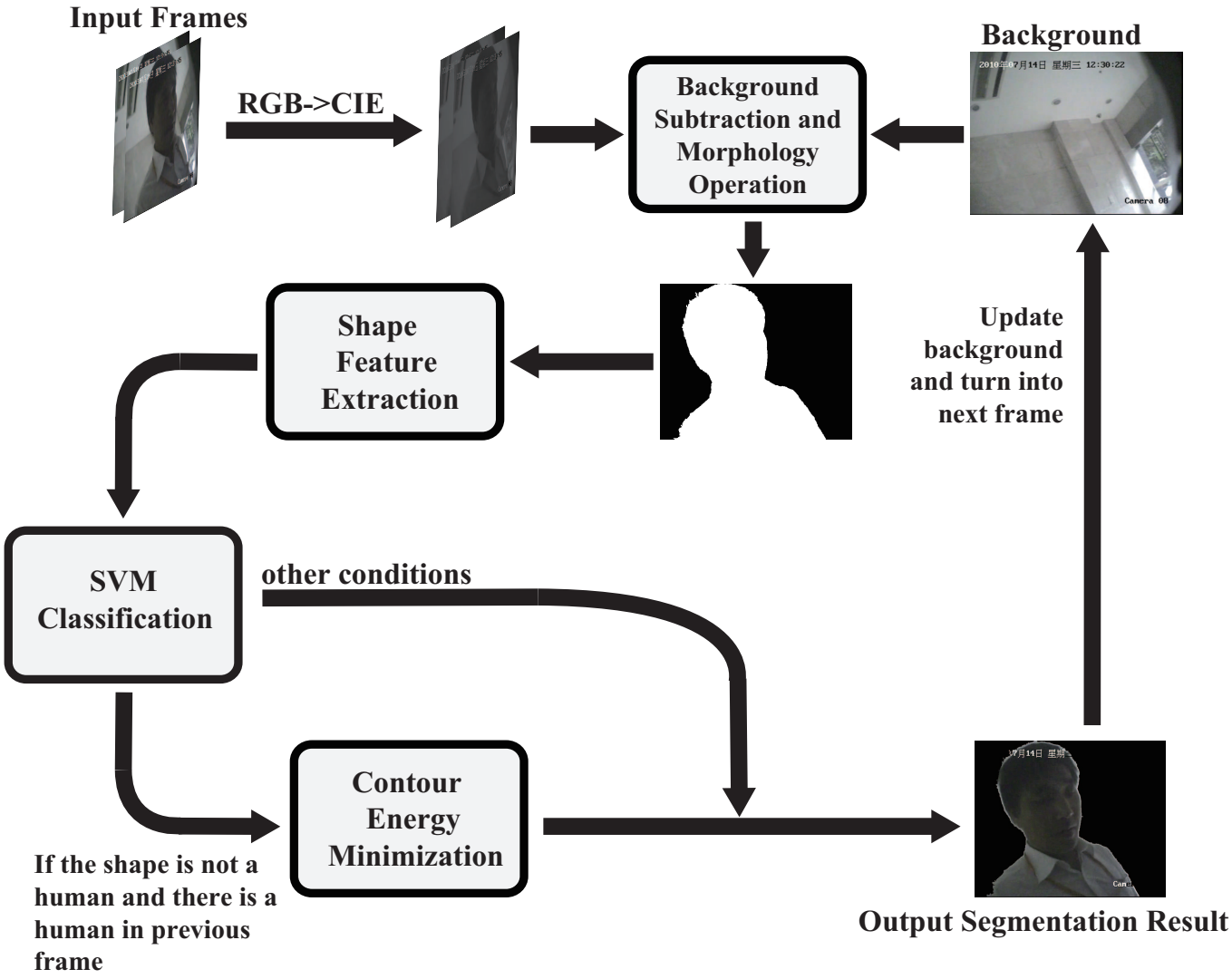


Fig. 1. The algorithm flow diagram of our method.

detection performance using non-overlapping multi-resolution HOG descriptors and a histogram intersection kernel SVM based on a spatial pyramid match kernel. These methods involve overlapped window-based feature computation which is time consuming. Thus they are difficult to be applied in videos.

Contrary to whole body, fewer researches are dedicated to the upper body part. Eichner et al. [13] developed an upper-body detector for near-frontal viewpoint based on HOG feature. However, it fails on low contrast videos because of the blurry boundary. Other proposed methods [14], [15] also used feature of human upper body. [14] represented the upper body as an assembly of four body parts. Detection of the body parts in single frame makes the method insensitive to camera motions. [15] used an offline boosted multi-view upper-body detector to automatically initialize a new human trajectory and it is capable of dealing with partial human occlusions. However, their edgelet feature which is local to the whole human body cannot completely describe a human without other inference mechanism.

B. Segmentation

Foreground segmentation, or background subtraction, has a longer history. Stauffer et al. [16] represented each pixel by a mixture of Gaussian distributions and updated each pixel with new Gaussians. Kim et al. [17] expressed background as cylindrical codebook model and updated the model adaptively. Sample background values at each pixel are quantized into codebooks which represent a compressed form of background model for a long image sequence. Liu et al. [18] sampled pixels along the time axis and clustered them with mean shift technique. Each cluster was assigned with a weight to demonstrate its likelihood of being background. Though these methods can handle illumination variation, they still fail once foreground objects keep still for a long time.

Human segmentation methods are usually combined with human detection. Gao et al. [19] proposed a new feature called ACF in oriented granular space for both detection and segmentation. This feature consists of a chain of granules in oriented granular space (OGS) that is learnt via the AdaBoost algorithm. Three operations are defined on the OGS

to mine object contour feature and feature co-occurrences automatically. A heuristic learning algorithm is proposed to generate an ACF that defines a weak classifier for human detection or segmentation. Lin et al. [20] used hierarchical part-template matching to segment human. Zhao et al. [21], [22] applied a Markov Chain Monte Carlo approach based on the knowledge of various aspects including human shape, human height, camera model, and image cues. These methods have the limitation that they cannot perform well on low contrast images and videos.

III. Main Idea and Algorithm Overview

Our aim is to extract a complete human upper body region in low contrast videos with illumination variation at a real-time rate. The difficulties include how to determine a human region in frames and how to find the correct human regions when illumination has changed while foreground keeps still for a long interval. Existing methods cannot be applied to address this problem because of its particularity.

Fig. 1 shows the complete algorithm we propose in this paper. We take the first frame as initial background. Input frames are converted from RGB color space to CIELab and subtract from the background to generate a coarse mask of foreground regions. Then connected components searching and morphology operations are applied on the mask to obtain refined foreground regions. Subsequently we collect sample points on the contour of a foreground region and map them onto a 2D histogram from which shape feature vectors are abstracted. A support vector machine classifier is in charge of identifying whether the input vectors belong to human or non-human based on the model pre-trained by positive and negative samples. Every independent foreground region is classified and regions identified as humans are recorded for updating background. The energy minimization procedure is not executed until a recorded region cannot be identified by the SVM classifier, for example, when the foreground expands drastically. Initialization of the optimizing procedure is the contour of identical region in last frame. The region surrounded by the output contour will be the rectified foreground and we update the background region to prevent segmentation error in subsequent frames.

IV. Upper Body Model Training and Detection with Shape Feature

In this section, a detailed demonstration about our shape feature to distinguish human and non-human will be shown. We also give a full specification of the training and detection procedures.

A. Shape Feature Description

One of specific features of a person is his/her shape which actually is a region encompassed by a contour in videos and images. Unlike HOG features [7] and edgelet features [14] which describe shape locally, our feature captures the entire shape of human while keeping the simplicity, thus it has more distinguishability and less computational complexity.

A person's shape, especially upper body shape, is assumed to be a star convex [23]. A region $\mathcal{C}$ is called a star convex if there exists a point x_0 in it such that the line segment from x_0 to any point y in $\mathcal{C}$ is contained in $\mathcal{C}$. That a region is a star convex means it is relatively simple, which is shown expressly in Fig. 2. From the figure one can conclude that all human's upper body shapes have a common shape pattern distinctive from other objects. Our feature is originated from this idea. For a certain foreground region, we find its centroid through

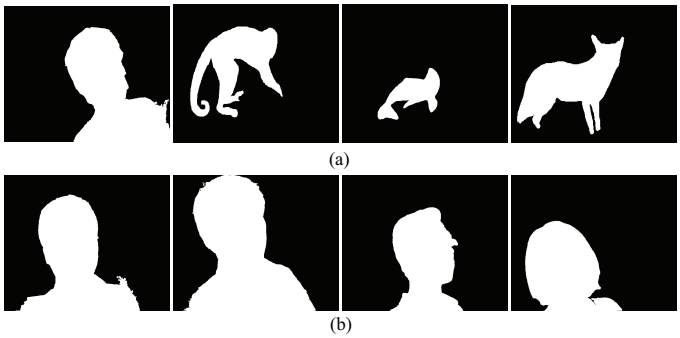


Fig. 2. Object shape diagram. (a) Human upper body shape has significant difference from other objects in 2D planar and is simpler and more regular than other creatures and objects. (b) Meanwhile all the human, male or female, have common shape features.

breadth-first based connected components searching algorithm and seek the contour of the region through a border following algorithm [24]. Unlike common sample strategies, our method samples points from the contour with even degree instead of even contour length. In other words, we construct a polar coordination with the region centroid as its origin and sample the points along counterclockwise direction. Initial sample point is specified as the one has the minimal y coordination (Fig. 3). Then for each point $(\theta_i, r_i), i = 1, 2, ..., N$, where r_i is the distance from the region's centroid to the sample point, θ_i is the inclined angle between vertical direction and the direction consistent with r_i, N is the total number of sample points, we project it onto a 2D histogram with θ_i as x axis and r_i as y axis. θ_i and r_i are quantified into m and n bins, respectively. The magnitude of a bin of the histogram means the number of sample points fallen into the corresponding bin. The whole histogram represents a distribution of contour shape after it is normalized. In Fig. 4, one can see that a human upper body shape has a specific shape pattern with head part smaller and shoulder part broader which results in a different histogram with other objects.

After voting, we use an $m \times n$-dimension vector $\mathcal{F} = \{f_1, f_2, ..., f_{m\times n}\}$ to represent the histogram (5 r-bins $\times$ 12 θ-bins in our implementation, so it forms a vector in a 60-dimension feature space). Every f_k in $\mathcal{F}$ is the normalized magnitude of the corresponding bin.

Obviously our shape feature has no relation with object's position in a frame. However, we do not consider rotation invariance since we assume all the humans in videos we process are upright, which is a common phenomenon in practice.

The differences between the new descriptor and the original

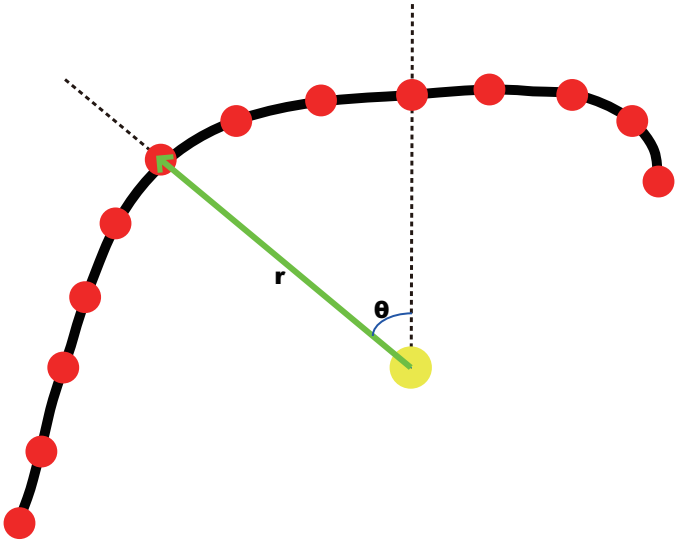


Fig. 3. Diagram of shape feature sample strategy.

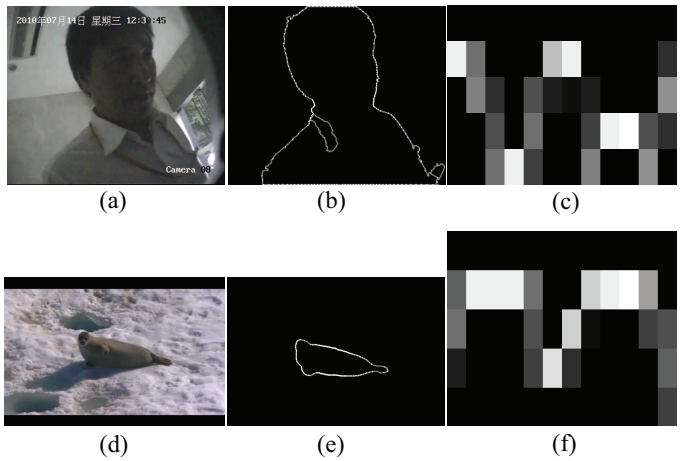


Fig. 4. Shape feature and its representation. (a)-(c) are a man's image, his shape feature and generated 2D histogram, respectively; (d)-(f) are the comparative ones of a seal.

Shape Context are the coordinate of the central position of the histogram and the sample strategy. Each point is sampled from the contour with even degree instead of even contour length. For Shape Context, each sample point will be considered with other points, which leads to $O(k^2)$ operations to generate descriptor for k points, while our method only needs $O(k)$. Moreover, since our descriptor has more concise information, it saves much time for computation in the classification phase.

B. Training and Detection

In training phase, we collect human and non-human images and manually extract foreground regions. Then we compute the shape features using the method in Section IV and feed them into a C-SVC (C-Support Vector Classification) classifier to train a hyperplane which can partition the high dimension feature space properly. In addition, we choose GRBF (Gaussian Radial Basis Function) as the kernel function of nonlinear SVM:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2) \quad (1)$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are feature vectors and γ is a normalized constant.

In detection phase, for an input frame we convert its color space from RGB to CIELab and set the first frame as background. We take the difference between current frame and background as the preliminary foreground regions. Then we use dilate morphological operation to eliminate small regions after connected components searching on foreground and background respectively. Taking generated foreground mask as input, the classifier obtained in training phase is used for outputting class labels 1 or -1. When illumination condition varies and makes the foreground change its shape, which results in the classifier taking a human as a non-human, we apply a contour evolution technique shown in section V. Finally, background frame is updated by a linear interpolation strategy:

$$I_B(x, y) = \alpha I_B^{(t)}(x, y) + (1 - \alpha) I_B^*(x, y) \quad (2)$$

where $I_B(x, y)$ is a pixel value of background frame on location (x, y) after updating and $I_B^*(x, y)$ is the the same frame pixel value before updating. $I_B^{(t)}(x, y)$ is the value of a pixel that belongs to a background region in current frame. α is a constant to control the updating speed. For pixels on the foreground location, we simply set $I_B(x, y) = I_B^*(x, y)$.

V. Contour Evolution Technique

A. Implementation

Once detected human regions cannot be discerned by the shape classifier in the current frame, we immediately start a contour evolution procedure. Assign an energy function $E(\mathbf{v}(s))$ to the contour:

$$E(\mathbf{v}(s)) = \oint (E_{int}(\mathbf{v}(s)) + \eta(s) E_{ext}(\mathbf{v}(s))) ds \quad (3)$$

where $E_{int}(\mathbf{v}(s))$ is internal potential energy of the contour and $E_{ext}(\mathbf{v}(s))$ is external image-based constraint forces [9]. $\eta(s)$ is a weight for every sample point, which is defined as:

$$\eta(s_i) = \frac{\|\nabla \mathbf{I}(x(s_i), y(s_i))\|^2}{\sum_i^N \|\nabla \mathbf{I}(x(s_i), y(s_i))\|^2} \quad (4)$$

where $\nabla \mathbf{I}$ refers to the gradient of the image, and N is the total number of sample points. Our aim is to find the function $\mathbf{v}(s) = (x(s), y(s))$ to minimize the energy functional. Euler-Lagrange method is applied to solve this problem [25]. Discretization of the two partial differential equations results in a matrix multiplication form $\mathbf{Ax} = \mathbf{b}$ where $\mathbf{A}$ is a pentadiagonal matrix. Cholesky decomposition can be applied to solve this linear sparse system.

B. Discussion

As in Fig. 1, the contour evolution is executed only when the classifier detects a human upper body shape in previous frame and a non-human upper body shape in current frame, which is caused by the illumination variation. And illumination

variation occurs infrequently with respect to the video length (about 3-4 times in a 5 minutes video). In most cases the video is processed by shape descriptor generation step and background updating step, which both have high efficiency. So the delay resulted from the time-consuming contour evolution algorithm can be compensated by fast descriptor generation and background updating.

Moreover, the initial contour in current frame used for evolution algorithm is the one in previous frame. Since the contour in previous frame has right shape and the displacement of foreground human body between two frames is small, it is obvious that we have a good initial value and a relatively fast convergence rate. Fig. 5 demonstrates the strategy.

(a) (b) (c)

Fig. 5. The contour evolution strategy. (a) The contour in previous frame. (b) When detecting non-human upper body contour, we take the contour in previous frame as the initial contour of the evolution process instead of current contour. (c) The convergence contour after contour evolution.

VI. Experimental Results

In this section we show several experimental results about human detection and segmentation. Meanwhile we compare performance and efficiency with other algorithms. Our experiments are carried out in following hardware environment: Core2 Duo 2.83GHz CPU, 4G RAM memory and GeForce GTX 460.

A. Detection Result

In this experiment we show the discriminability of our shape feature. There is no available dataset of human's upper body, so we collect images of 500 human and non-human samples from our ATM videos and internet, then manually extract their shape regions represented by mask.

To train our human classifier, we must determine the classifier's parameters, in SVM case they are γ for the kernel function and C for the cost, to achieve the best classification performance. In this experiment, a n-folds validation is carried out: i.e., $(n-1)/n$ of each class is selected as training datasets while the remaining as testing datasets. This process is repeated n times with different combinations of training and testing datasets, and the combination has the highest classification accuracy is chosen. According to the cross validation (CV) contour shown in Fig. 6, the best combination is $\gamma = 0.25$ and $C = 2.0$.

We test our classifier on several representative ATM videos from real world. And these test sequences are different from the training set. The resulting precision and recall are shown in Table I. We test the HOG based human upper body detector [13] on the same videos (the results are also shown in Table I).

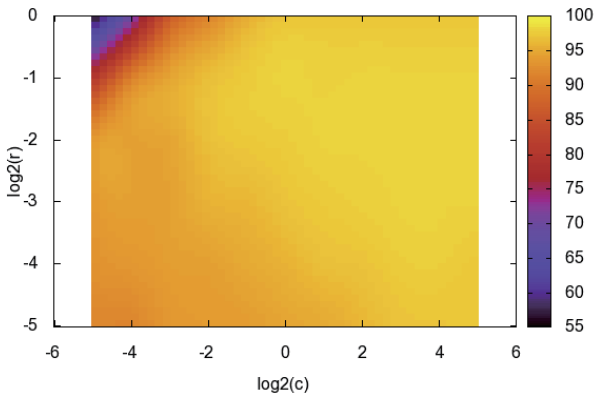


Fig. 6. Different classification accuracies represented by the cross validation contour. The range of γ is $[2^{-5}, 1]$ and the one of C is $[2^{-5}, 2^5]$. For simplicity, we plot them with their logarithms. Different color represents different accuracy. We can see that the accuracy of most combinations of C and γ is above 90% except the left top corner part. This demonstrates that our shape feature is robust and reliable.

TABLE I
THE COMPARISON OF PRECISION AND RECALL RESULTS ON SEVERAL ATM VIDEOS BETWEEN OUR PROPOSED METHOD AND [13]. IN THIS TABLE, TF MEANS THE NUMBER OF TOTAL FRAMES. FOH IS FRAME NUMBER OF HUMAN. TP, FP AND FN REPRESENT TRUE POSITIVE, FALSE POSITIVE AND FALSE NEGATIVE, RESPECTIVELY. THE PRECISION (PREC) IS COMPUTED AS $TP/(TP+FP)$ AND RECALL (REC) $TP/(TP+FN)$.

Ours	TF	FoH	TP	FP	FN	Prec(%)	Rec(%)
Video1	6292	6105	5890	8	215	99.9	96.5
Video2	1250	764	749	20	15	97.4	98
Video3	3750	3170	3092	24	78	99.2	97.5
Video4	2996	2576	2496	19	80	99.2	96.9
Video5	2533	1620	1592	33	28	98	98.3
[13]	TF	FoH	TP	FP	FN	Prec(%)	Rec(%)
Video1	6292	6105	2430	6	3675	99.8	39.8
Video2	1250	764	14	0	750	100	1.8
Video3	3750	3170	1866	2	1304	99.9	58.9
Video4	2996	2576	539	1	2037	99.8	20.9
Video5	2533	1620	774	0	846	100	47.8

From Table I we can see that both methods have comparative precision but our proposed method has much higher recall. For a fair comparison, we trained [13] on our ATM image data. For several sequences, such as those recorded in the daytime, the HOG based detector [13] can locate most upper bodies. However, for sequences taken in night, it dose not work because the low contrast leads to small gradient on the boundary which is fatal to HOG feature. Moreover, the HOG feature is more sensitive than our proposed feature to the shape. For example, when a part of a man or woman's upper body is truncated since the limited view of cameras, the incomplete gradient information generate an imperfect HOG

feature which increases the false negatives rate. Fig. 7 shows the scenes that HOG detector cannot handle.

Meanwhile we compare the efficiency of the two methods. Because of HOG feature extraction step's low efficiency and larger foreground area in the ATM application, the HOG-based detector is much slower than our method. For test convenience, we downsample the resolution of original frame from $704*576$ to $352 * 288$. The results are shown in Table II.

TABLE II
THE TIME COMPARISON (IN MS, AVERAGED PER FRAME) BETWEEN [13] AND OUR METHOD.

	Ours	[13]
Video1	15.0	1054
Video2	16.8	1288
Video3	21.3	1457
Video4	17.3	961
Video5	16.6	1755

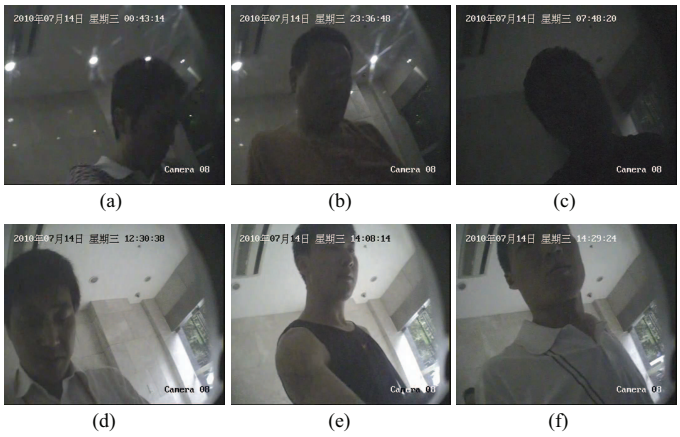


Fig. 7. The situations that [13] cannot handle. (a)-(c) Low contrast makes human upper body's boundary fuzzy. (d)-(f) Parts of the human body are truncated which provide incomplete gradient information.

B. Segmentation Result

In this experiment we compare our segmentation method with several popular foreground object extraction algorithms. From experimental results we can see that other algorithms either mistake foreground regions as background or misjudge background regions as foreground(Fig. 8). The reason is that people in these videos keep still for a relatively long time so gradually they are treated as background by traditional background subtraction methods. On the contrary, our method combine background subtraction, shape feature classification and active contour to achieve a better performance.

Our proposed method can accomplishes a real-time speed while keeping the segmentation results acceptable. The shape feature proposed in our paper can classify a shape, which is represented by a mask, as a human or non-human. The mask is generated by background subtraction. When illumination variation happens, the shape of generated mask also changes,

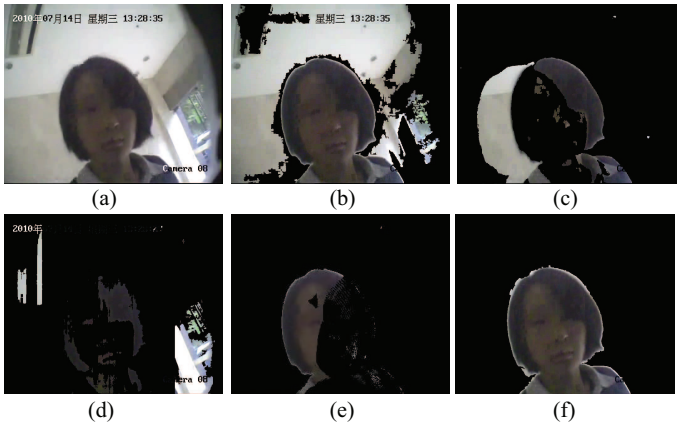


Fig. 8. (a) Original frame of a test video. (b)-(f) The segmentation results: (b) Background subtraction: modeling each pixel in background as a single Gaussian model and for each frame subtracting it from the background to get the foreground; (c) The algorithm of [16]; (d) The algorithm of [18]; (e) The algorithm of [17]; (f) Our proposed method.

which will cause it be classified as non-human. While in the previous frame, it was classified as a human's upper body. This means the currently detected contour is wrong because it conflicts with temporal coherence, as shown in Fig. 9. Once it happens, we will correct the wrong contour by evolution. Especially, the initial contour in current frame used for evolution algorithm is the one in previous frame. Since the contour in previous frame has right shape and the displacement of foreground human body between two frames is small, it is obvious that we have a good initial value and a relatively fast convergence rate. The shape feature generation and classification steps require only 2 ms averagely. The most time-consuming step is energy function minimization which takes average 182 ms. Since this step is applied only when foreground regions are classified mistakenly due to illumination variation, the delay at these special frames can be recovered by the rapid processing at normal frames. Fig. 10 shows more segmentation results of our methods.

Fig. 9. The generated mask before and after illumination variation. (a) The mask before illumination variation; (b) The mask after illumination variation.

VII. CONCLUSION

In this paper we propose a novel method for human detection and segmentation. It consists of two main procedures: the first polar coordination based shape feature generation

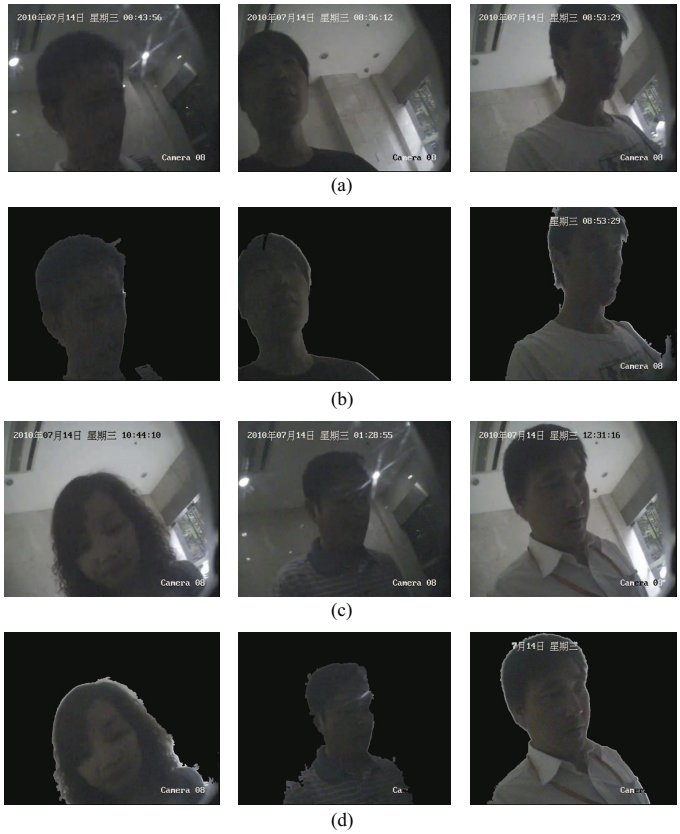


Fig. 10. More segmentation results of our proposed method. (a) and (c) are original input videos while (b) and (d) are corresponding segmentation results, respectively. We can see that our proposed method can handle video taken at night and with incomplete view except for normal conditions.

and SVM classification; and the second a contour energy minimization procedure. The two procedures mutually improve with each other and accomplish a good performance. In experiments we show the acceptable segmentation results in changing illumination environment as well as high detection rate and real-time performance. The detection range of body dimension is from the part above human's neck to the part above human's chest due to the training set obtained from ATM's sequences. In the future, we will try to use Salient Region Detection [27] and Human Pose Database [28] to further improve our method.

Our proposed method cannot be used on static images and currently it can only detect one object. This is a limitation of our method.

ACKNOWLEDGMENTS

The authors would like to thank C.C. Chang and C.J. Lin for providing libsvm and M. Eichner etc. for providing human upper body detector on their website.

REFERENCES

[1] D. Lee, P. Zhan, A. Thomas, and R. Schoenberger, "Shape-based human intrusion detection", in *SPIE Int. Symposium on Defense and Security, Visual Information Processing XIII*, vol. 5438, pp. 81-91, Apr. 2004.

[2] J. Zhou, and J. Hoang, "Real time robust human detection and tracking system", in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*,vol. 3, pp. 149-156, Jun. 2005.

[3] D. Toth and T. Aach, "Detection and recognition of moving objects using statistical motion detection and fourier descriptors", in *Int. Conf. on Image Analysis and Processing*, pp. 430-435, Sep. 2005.

[4] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: real-time tracking of the human body", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, pp. 780-785, Jul. 1997.

[5] H. Eng, J. Wang, A. Kam, and W. Yau, "A bayesian framework for robust human detection and occlusion handling using a human shape model", in *Int. Conf. on Pattern Recognition*, vol. 2, pp. 257-260, Aug. 2004.

[6] H. Elzein, S. Lakshmanan, and P. Watta, "A motion and shapebased pedestrian detection algorithm", in *IEEE Intelligent Vehicles Symposium*, pp. 500-504, Jun. 2003.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, pp. 886-893, Jun. 2005.

[8] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 509-522, Apr. 2002.

[9] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models", *International Journal of Computer Vision*, vol. 1, pp. 321-331, Jan. 1988.

[10] P. Viola, M. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance", in *IEEE Int. Conf. on Computer Vision*, vol. 2, pp. 734-741, Oct. 2003.

[11] C. Hou, H. Ai, and S. Lao, "Multiview pedestrian detection based on vector boosting", in *Asian Conf. on Computer Vision*, vol. 4843, pp. 210-219, Nov. 2007.

[12] S. Maji, A. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient", in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp. 1-8, Jun. 2008.

[13] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "Articulated human pose estimation and search in (almost) unconstrained still images", D-ITET, BIWI, ETH Zurich, Tech. Rep. No.272, Sep. 2010.

[14] B. Wu, and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors", *Int. Journal of Computer Vision*, vol. 75, pp. 247-266, Nov. 2007.

[15] J. Xing, H. Ai, and S. Lao, "Multiple human tracking based on multi-view upper-body detection and discriminative learning", in *Int. Conf. on Pattern Recognition*, pp. 1698-1701, Aug. 2010.

[16] C. Stauffer, and W. Grimson, "Adaptive background mixture models for real-time tracking", in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, pp. 246-252, Jun. 1999.

[17] K. Kim, T. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground-background segmentation using codebook model", *Journal of Real-Time Imaging*, vol. 11, pp. 172-185, Jun. 2005.

[18] Y. Liu, H. Yao, W. Gao, X. Chen, and D. Zhao, "Nonparametric background generation", in *Int. Conf. on Pattern Recongnit.*, vol. 4, pp. 916-919, Sep. 2006.

[19] W. Gao, H. Ai and S. Lao, "Adaptive contour features in oriented granular space for human detection and segmentation", in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp. 1786-1793, Jun. 2009.

[20] Z. Lin, and L. Davis, "A pose-invariant descriptor for human detection and segmentation", in *Proc. IEEE European Conf. on Computer Vision*, vol. 5305, pp. 423-436, Oct. 2008.

[21] T. Zhao, and R. Nevatia, "Bayesian human segmentation in croweded situations", in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, pp. 459-466, Jun. 2003.

[22] T. Zhao, and R. Nevatia, "Stochastic human segmentation from a static camera", in *Workshop on Motion and Video Computing*, pp. 9-14, Dec. 2002.

[23] G. Rawlins, and D. Wood, "Ortho-convexity and its generalizations", *Computational Morphology*, pp. 137-152, 1988.

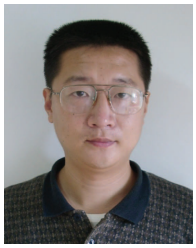
[24] S. Suzuki, and K. Abe, "Topological structural analysis of digitized binary images by border following", *Computer Vision, Graphics, and Image Processing*, vol. 30, pp. 32-46, Apr. 1985.

[25] F. Leymarie, and M. Levine, "Tracking deformable objects in the plane using an active contour model", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, pp. 617-634, Jun. 1993.

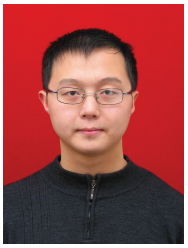
[26] C. Chang, and C. Lin, "LIBSVM: a library for support vector machines", 2001.

[27] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection", in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp. 409-416, Jun. 2011.

[28] T. Chen, P. Tan, L. Ma, M. Cheng, A. Shamir, S. Hu, "PoseShop: Human Image Database Construction and Personalized Content Synthesis", *IEEE Trans. on Vis. Comput. Graph.*, vol. 99, PrePrints, 2012.



Ruofeng Tong received the BS degree in 1991 in the Department of Mathematics at Fudan University and PhD in 1996 in the Department of Mathematics at Zhejiang University, China. Currently, he is a Professor in the College of Computer Science and Engineering, Zhejiang University. His research interests include computer vision, image processing and CAD&CG.



Di Xie is a Ph.D. candidate in the Department of Computer Science, Zhejiang University, China. He received his BS degree from Zhejiang University, China, in 2007. His research interests include image and video processing and computer vision.

Min Tang received the BS degree in 1994 and Ph.D degree in 1999 at the Department of Computer Science and Engineering from Zhejiang University, China. Currently, he is a Professor at Zhejiang University. He was a visiting scholar at the Department of Computer Science, Wichita State University in 2006, and at the Department of Computer Science, UNC-Chapel Hill in 2008. His research interests include image processing, CAD&CG and collision detection.