REAL-TIME CLOTHING RECOGNITION IN SURVEILLANCE VIDEOS

Ming Yang, Kai Yu

Media Analytics Dept.
NEC Laboratories America, Inc.
Cupertino, CA 95014, USA

ABSTRACT

Recognition of clothing categories from videos is appealing to emerging applications such as intelligent customer profile analysis and computer-aided fashion design. This paper presents a complete system to tag clothing categories in real-time, which addresses some practical complications in surveillance videos. Specifically, we take advantage of face detection and tracking to locate human figures and develop an efficient clothing segmentation method utilizing Voronoi images to select seeds for region growing. We compare clothing representations combining color histograms and 3 different texture descriptors. Evaluated on a video dataset with 937 persons and 25441 cloth instances, the system demonstrates promising results in recognizing 8 clothing categories.

***Index Terms*—** Clothing Recognition, Cloth Segmentation, SVM

1. INTRODUCTION

Clothing recognition is an advanced image processing application, which may benefit customer profile analysis, context-aided people identification [1, 2, 3], and computer aided fashion design [4, 5, 6]. Although this problem attracts increasing research interests [4, 7, 8, 9, 3, 5, 6] in recent years, a real-time clothing recognition system, especially for surveillance videos, remains challenging, primarily due to two reasons. First, such a system involves a series of difficult sub-problems including face detection and tracking, human figure or clothing segmentation, and effective clothing representations. Second, the differences among various clothing categories are inherently subtle and even vague for human, thus considerable computations are required to discern them.

In this paper, we present a video content analysis system which is capable of tagging clothes of multiple persons to some pre-defined categories, *i.e.*, *suit*, *shirt*, *T-shirt*, *jeans*, *short pant*, *short skirt* and *long skirt*, in real-time, as illustrated in Figure 1. The main contributions include: 1) We develop a system design tailored for clothing analysis including an efficient clothing segmentation method; 2) We evaluate different clothing representations combining color histograms with histogram of oriented gradient (HOG), Bag-of-Words (BOW) features, and DCT responses. Experiments on a test set with 937 persons and 25441 cloth instances demonstrate the average recall over 80% at false positive rate 0.1. To our best knowledge, the research and development efforts that are dedicated to a practical solution to clothing recognition particularly for surveillance quality videos have not been reported before.

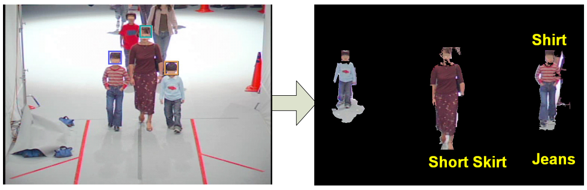

Fig. 1. Recognition of certain clothing categories from surveillance videos.

Digital analysis of clothing images can trace back to 90s. People first studied how to segment clothes as foreground objects [4, 7, 9, 3] assuming some foreground seeds [4] or regions of interests [7] provided by users, or the same clothes appearing in different backgrounds [3]. Later, rough clothing analysis was employed as contexts to help people identification [1, 2, 3]. A comprehensive high-level modeling of clothes based on And-Or graphs was presented in [8]. Responsive [5] and smart mirrors [6] proposed to retrieve similar clothing styles for fashion recommendation in a fitting room, which are the closest work to ours. However, in surveillance videos, people are not as cooperative as in a fitting room, thus, we have to take into consideration the complications such as scale changes and partial occlusions. Also, we assign clothing categories as tags without any query image. More importantly, most of the existing approaches focus on processing static images of a single person, hence, the computational costs tend to be too high for video analysis. In contrast, our system strives to keep the computations affordable for video analysis and processes multiple persons simultaneously in real-time.

2. SYSTEM DESIGN

Video content analysis systems confront by many complications in surveillance videos. For example, when the surveillance camera is mounted at the entrance of a shopping

978-1-4577-1303-3/11/$26.00 ©2011 IEEE

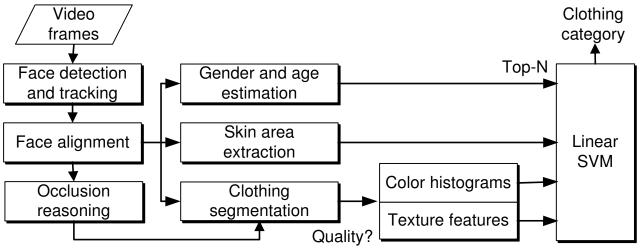

Fig. 2. Clothing recognition system diagram.

mall, the interaction of multiple persons may lead to partial occlusions and cluttered background, not to mention the scale and view angle changes. We address these difficulties by integrating diverse features from multiple cloth instances.

Clothing recognition involves several critical sub-tasks including localization of human figures, clothing segmentation and alignment, and extraction of clothing representation. None of them is a fully solved problem for surveillance videos. Hence, we cannot expect reliable results from every single frame. Thus, we collect cloth instances on the trajectory of a person and preserve the top N good instances which are measured by non-occluded areas and the quality of segmentation results, then the average features of these instances are used to represent clothes. Since clothing categories are high-level semantic concepts, it is desirable to integrate various relevant clues to recognize them, such as a person's gender and age, uncovered skin areas, color and texture features of clothes. In particular, it is crucial that these color and texture features are extracted from the image regions of clothes, not from background nor nearby persons.

The system diagram is summarized in Figure 2. First, we perform face detection and tracking [10] for each frame and align the detected faces. Then, we crop a candidate rectangular region based on each aligned face (5 times of the face width and 9 times of the face height). Simple occlusion reasoning among persons is conducted according to face locations and the overlapped areas of the candidate rectangles. For the persons with a visible frontal face and moderate occlusions, *i.e.*, the non-occluded area is larger than 75% of the cropped rectangle, we proceed to segment the clothing from this candidate region using an efficient variant of region growing method. A cloth is represented by the average feature vector of $N = 10$ instances, including his/her estimated gender and age [11], skin area ratios of arms and legs, 2D color histograms in the LAB space, and texture descriptors. Afterwards, multi-class linear SVM classifiers are employed to learn the clothing categories. To make our system general for different scenarios, we assume no user interaction in the process at all. The segmentation method and the clothing representation are described in details in Sec. 3 and Sec. 4, respectively.

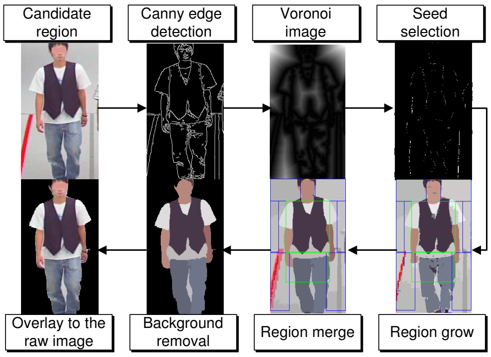

Fig. 3. The flow chart of clothing segmentation.

3. CLOTHING SEGMENTATION

Clothing segmentation and proper human figure alignment are the prerequisites to perform feature extraction and clothing classification, thus they are vital to the recognition accuracy. We develop a lightweight variant of the region growing based color segmentation method [12] for foreground clothing segmentation. Although the cameras are stationary for current test videos, in order to ensure the applicability of the system for moving cameras we do not utilize motion information or apply background subtraction technique.

Given the candidate rectangle region of a person, we segment it to roughly homogeneous color segments, then apply the prior knowledge of foreground and background based on the face alignment to extract the foreground figure or clothes. The procedure is illustrated in Figure 3.

We employ the Voronoi image [12] to automatically place the seed points close to the centers of homogeneous regions. The intensity at each point on a Voronoi image is the distance to the closest edge, hence, the peaks reflect the farthest points from the contours. Therefore, we conduct the Canny edge detection on gray-level images and choose the local maximums on the Vonoroi image as seeds for region growing. It is worth noting that by discarding small local maximums that are less than $T = 5$, too small segments are merged, so we mitigate the over-segmentation problem for texture regions to some extent. Next, we employ L2 norm of two color pixels in the perceptual uniform LAB color space to measure their discrepancy and obtain rough homogeneous color segments after region growing. The color pixels in one segment are modeled as a Gaussian distribution, then two adjacent segments are merged if the Mahalanobis distance of two color distributions is small. Small color segments are merged to the most similar segments next to them. Finally, we apply the prior knowledge of foreground (drawn as green rectangles in Figure 3) and background (drawn as blue rectangles in Figure 3) to extract the foreground clothing and measure the segmentation quality. Further the foreground mask is used to extract the color and texture features of clothing.

The advantage of this method is the computational efficiency which is critical for real-time applications. This region

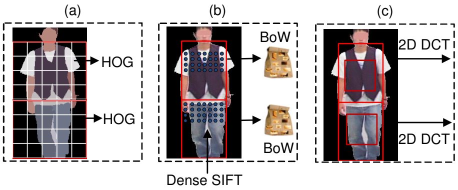

Fig. 4. Different texture features based on HOG, BoW, and DCT responses.

growing method takes about 15-18 ms to process candidate image regions with resolution 200 × 300. We have also implemented the normalized-cut based color segmentation [13] which takes nearly a second to process images with similar sizes. Yet the region growing method may suffer from the so-called leakage problem occasionally.

4. CLOTHING RECOGNITION

Since the clothing segmentation results may not be reliable for every single frame due to partial occlusions or the occasional leakage problem, we employ a few cloth instances with good segmentation quality to calculate the average feature vector to represent a cloth. The quality of clothing segmentation is measured by the difference between the ratios of overlapped areas of foreground clothes with the foreground priors and background priors in Figure 3. We keep a set of up to $N = 10$ cloth instances for a person. The features include the estimation of gender and age [11], the skin area ratios of limbs, color histograms, and texture features.

The skin areas of arms and legs uncovered by clothes are an informative clue for clothing category. Though the skin tone may appear distinct under different lighting conditions, the color on the face is usually consistent with color of arms and legs. We compute a 32 × 32 2D color histogram from the aligned frontal face in the LAB color space. The dominant color component is regarded as the skin color tone to label the skin pixels on arms and legs. The ratios of skin pixels over the areas of limbs generate a 4D feature vector.

We analyze the clothing categories of the top and bottom parts separately. The clothing segmentation result is used as a mask when calculating the color and texture features. The color of the clothes is delineated by a 2D color histogram in the LAB color space with 32 bins for each channel, that is a 1024D feature vector. We evaluate various texture features based on histogram of oriented gradient (HOG) in multiple spatial cells, a bag of dense SIFT features [14], and DCT responses. The gradient orientations on 8 directions (every 45 degree) are computed on the color segmentation results, not the original raw images. The top and bottom parts are both spatially divided to 4×5 cells, drawn as white rectangles in Figure 4a, in which the histograms of all cells are concatenated to an $8 \times 20 = 160D$ HOG feature. The dense 128-dimension SIFT features are calculated every 6 pixels inside the top and bottom parts of a human body which are drawn in the red rectangles in Figure 4b. These local features are quantized with a visual codebook with 256 words. The frequencies of the codewords are normalized to generate the BoW descriptor. As illustrated in Figure 4c, the regions in the two red rectangle cells are resized to 16×16 patches, then the first 128 components of the DCT coefficients in the zigzag scan are packed to form a $128 \times 2 = 256D$ DCT feature.

Table 1. The statistics of the test set.

Category	# of persons	# of instances
Suit (top)	47	1636
Suit (bottom)	47	1636
Shirt	100	2649
T-shirt	86	2314
Jeans	164	4880
Short pant	23	499
Short skirt	153	4182
Long skirt	44	1326
Total	937	25441

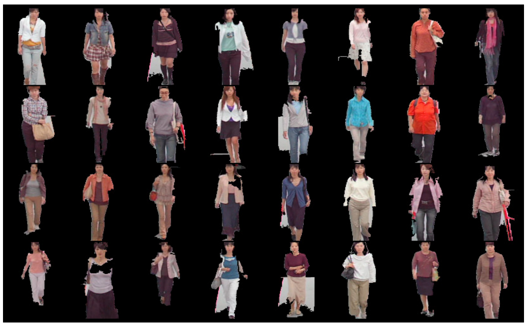

Fig. 5. Example clothing segmentation results.

The gender and age, the ratios of skin areas, color histograms, and texture features are concatenated as the clothing representation. A one-against-all linear SVM classifier is learnt for each clothing category. We have compared these 3 texture features and their combination. From the evaluation, the HOG feature outperforms the other two, which implies the shapes of color segments reveal more information about clothing categories than the textures of original raw images.

5. EXPERIMENTS

We collect the test video set assuming the camera is mounted at a mall entrance, which includes 937 persons and 25441 cloth instances obtained by face detection and tracking. We manually label 8 categories: *suit (top)*, *suit (bottom)*, *shirt*, *T-shirt*, *jeans*, *short pant*, *short skirt* and *long skirt*. The statistics are shown in Table 1. A sample frame is shown in the left of Figure 1. Note some clothes cannot be categorized to any one of these 8 classes. This is the largest clothing dataset reported in literature.

We employ a detection-driven approach [10] to locate and track faces. The face alignment and gender and age estimation modules are based on convolutional neural networks [11]. For videos at resolution 720 × 480, segmentation, feature extraction and classification run at 16-20 fps, and the entire system including face detection and tracking runs at 10

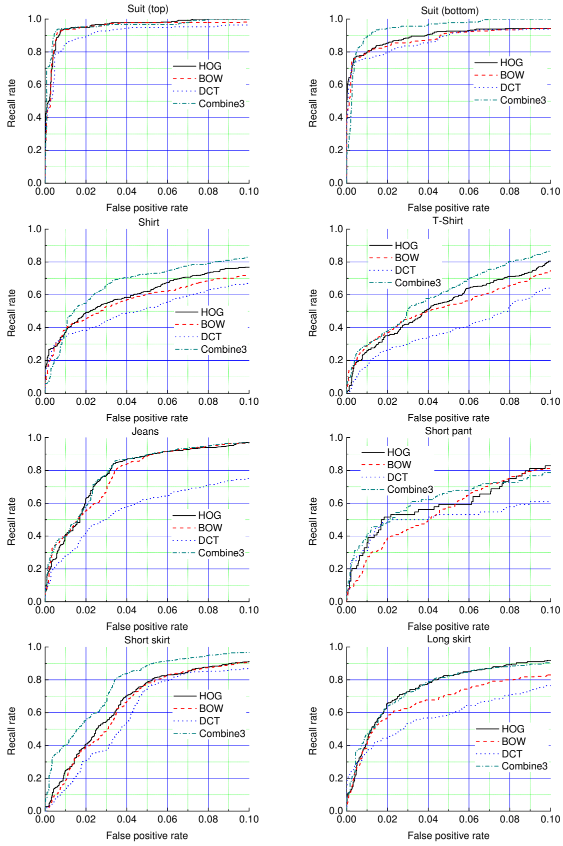

Fig. 6. ROC curves for 8 clothing categories up to FPR 0.1.

fps on an Intel Core 2 Duo 3.16Ghz desktop. The system throughput increases when down-scaling video frames since face detection is the computational bottleneck. Representative clothing segmentation results are shown in Figure 5 [1], where we observe for most cases that the efficient region growing method can yield acceptable foreground clothes.

We compare the clothing representations using the 3 texture features and their combination, denoted by *HOG*, *BoW*, *DCT*, and *Combine3*. The recognition performance is measured by an instance-based 5-fold cross-validation. The receiver operating characteristic (ROC) curves of the 8 categories up to false positive rate 0.1 are shown in Figure 6. The *HOG* feature consistently outperforms the other two texture features on all the 8 categories and shows very similar performance as *Combine3* for *suit (top)*, *jeans*, and *long skirt*, which suggests the shape and texture of the color segments are more informative regarding to clothing categories. The average recall and precision rates of *Combine3* at false positive rate (FPR) 0.01 are shown in Table 2. The performance varies among different categories. At FPR$= 0.01$, the recall of *suit* is over 90% with precision rate about 85%, while, for *T-shirt*, the recall rate is about 30% with the precision 70%. At FPR$= 0.1$, the recall rates exceed 80% for all 8 categories.

[1] The privacy of the persons must be protected to the maximum extent.

Table 2. The recognition performance of combining all 3 texture features at FPR 0.01.

Category	recall	precision
Suit (top)	94.2%	87.5%
Suit (bottom)	89.1%	85.7%
Shirt	37.9%	81.8%
T-shirt	29.1%	70.7%
Jeans	39.5%	90.1%
Short pant	39.8%	45.0%
Short skirt	40.2%	90.3%
Long skirt	46.7%	74.7%

Though not directly comparable, the performance is higher than that reported in [6].

6. CONCLUSIONS

This paper presents a practical clothing recognition system for surveillance videos. We have developed an efficient color segmentation method and thoroughly studied 3 types of texture features tailored for real-time recognition of clothing categories. The system achieves average recall rate 80% at FPR 0.1 in tagging 8 clothing categories.

7. REFERENCES

[1] Y. Song and T. Leung, “Context-aided human recognition - clustering,” in *Proc. ECCV*, May 2006, pp. 382–395.

[2] D. Anguelov, K. C. Lee, S. B. Gokturk, and B. Sumengen, “Contextual identity recognition in personal photo albums,” in *Proc. CVPR*, 2007.

[3] A. C. Gallagher and T. Chen, “Clothing cosegmentation for recognizing people,” in *Proc. CVPR*, June 2008.

[4] C. C Chang and L. L Wang, “Color texture segmentation for clothing in a computer-aided fashion design system,” *Image and Vision Computing*, vol. 14, no. 9, pp. 685–702, October 1996.

[5] W. Zhang, T. Matsumoto, J. Liu, M. Chu, and B. Begole, “An intelligent fitting room using multi-camera perception,” in *Proc. Int. Conf. Intelligent User Interfaces*, Janurary 2008, pp. 60–69.

[6] X. Chao, M. J. Huiskes, T. Gritti, and C. Ciuhu, “A framework for robust feature selection for real-time fashion style recommendation,” in *Proc. ACM Multimedia Workshop Interactive Multimedia for Consumer Electronics*, October 2009, pp. 35–42.

[7] A. Borras, F. Tous, J. Llados, and M. Vanrell, “High-level clothes description based on colour-texture and structural features,” in *Iberian Conf. Pattern Recognition and Image Analysis*, 2003, pp. 108–116.

[8] H. Chen, Z. J Xu, Q. Liu, and S. C Zhu, “Composite template for cloth modeling and sketching,” in *Proc. CVPR*, June 2006.

[9] Z. Hu, H. Yan, and X. Lin, “Clothing segmentation using foreground and background estimation based on the constrained delaunay triangulation,” *Pattern Recognition*, vol. 41, no. 5, pp. 1581–1592, May 2008.

[10] M. Yang, F. Lv, W. Xu, and Y. Gong, “Detection driven adaptive multi-cue integration for multiple human tracking,” in *Proc. ICCV*, Sept.29 - Oct.2, 2009, pp. 1554–1561.

[11] M. Yang, S. Zhu, F. Lv, and K. Yu, “Correspondence driven adaptation for human profile recognition,” in *Proc. CVPR*, June 2011.

[12] J. Freixenet, X. Munoz, D. Raba, J. Marti, and X. Cufi, “Yet another survey on image segmentation: Region and boundary information integration,” in *Proc. ECCV*, May 2002, pp. 408–422.

[13] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[14] D.G. Lowe, “Distinctive image features from scale invariant keypoints,” *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.