MULTI-CUE BASED CROWD SEGMENTATION

Ya-Li Hou and Grantham K. H. Pang
Department of Electrical and Electronic Engineering, The University of Hong Kong
Pokfulam Road, Pokfulam, Hong Kong

Keywords: Crowd segmentation, Human detection, Occlusions, Coherent motion.

Abstract: With a rough foreground region, crowd segmentation is an efficient way for human detection in dense scenarios. However, most previous work on crowd segmentation considers shape and motion cues independently. In this paper, a method to use both shape and motion cues simultaneously for crowd segmentation in dense scenarios is introduced. Some results have been shown to illustrate the improvements when multi-cue is considered. The contribution of the paper is two-fold. First, coherent motion in each individual is combined with shape cues to help segment the foreground area into individuals. Secondly, the rigid body motion in human upper-parts is observed and also used for more accurate human detection.

1 INTRODUCTION

Human shape is an important feature for human detection. However, shape cues may get less reliable when the background is cluttered or the crowd density is high. Motion features are usually examined in two categories. Some methods analyze the motion difference between two consecutive frames. Others find some motion characteristics based on multiple consecutive frames, like periodic motion, coherent moving trajectories. Our target in this paper is to develop an efficient method for crowd segmentation using cues from both shape and motion simultaneously.

2 RELATED WORK

People counting and human detection has become a hot topic in these years. All the methods based on shape cues can be classified into two categories. Their extension with motion features will also be discussed.

The first category exhaustively searches an image with a sliding window. Each window is classified as human or non-human with an advanced classifier based on shape, color or texture features (Dalal and Triggs, 2005, Zhu et al., 2006, Tuzel et al., 2007, Wang et al., 2009). These methods are usually extended by considering motion between two consecutive frames, as (Viola et al., 2003, Dalal et al., 2006). However, the methods in this category are computationally expensive. Lin et al. (Lin et al., 2007) use a template matching method to detect individuals in the crowd. A hierarchy of templates is established to include as many postures as possible since accurate template model are necessary in template matching-based methods. This method has faster computation speed. However, it is not straightforward to extend the method with motion features.

The other category assumes that a foreground area for the crowd has been obtained. People counting and detection are achieved by segmenting the foreground into individuals, like (Zhao and Nevatia, 2004, Zhao et al., 2008, Rittscher et al., 2005, Hou and Pang, 2009). Zhao and Nevatia (Zhao and Nevatia, 2004) locate the individuals in the foreground area by head detection. Head candidates are detected by checking local peaks on the foreground contour. A detected individual is removed from the foreground and the next individuals are detected in the remaining foreground region. Rittscher et al. (Rittscher et al., 2005) tried to reduce the requirements for an accurate foreground contour by only sampling some informative feature points on the contour. A variant of EM (Expectation-Maximization) algorithm is used to find the best grouping of the points with rectangles. Except for the background subtraction, the methods do not consider the motion features explicitly.

Recently, Hou and Pang (Hou and Pang, 2010) proposed a method for crowd segmentation in a

Hou Y. and K. H. Pang G..
MULTI-CUE BASED CROWD SEGMENTATION.
DOI: 10.5220/0003528901730178
In *Proceedings of the 8th International Conference on Informatics in Control, Automation and Robotics* (ICINCO-2011), pages 173-178
ISBN: 978-989-8425-75-1

dense crowd. They used the B-ISM (Block-based Implicit Shape Model) to explore the shape cues in the crowd, which can handle the ambiguity inside the dense area and reduce the requirements for accurate foreground contour. In this method, crowd segmentation is formulated as a feature point clustering process, which provides a nice framework to combine the shape cues with multi-frame motion features as mentioned in (Brostow and Cipolla, 2006, Rabaud and Belongie, 2006). In Brostow and Cipolla (Brostow and Cipolla, 2006), Rabaud and Belongie (Rabaud and Belongie, 2006), the motion characteristics of a person in multiple consecutive frames are observed. It is shown that interest points from an individual would display consistent trajectories while points from different persons usually have different trajectories. Their results in very crowded scenarios have shown the potential use of this idea for crowd segmentation. However, false alarms are quite likely to occur in the method when pedestrian exhibit sustained articulations. Also, very little shape information has been explored in these methods. As far as we know, there has been little work on combining multi-frame motion features with a shape-based method.

3 THE DEVELOPED METHOD

The method includes two stages: training stage and testing stage. Our focus is on the combination of the motion and shape cues for crowd segmentation in this paper. The details of the method will be described in this section.

3.1 Training Stage

In the training images, selected persons are annotated with a rectangle and the foreground region for training images is available.

First, interest points are detected on the training images and most points from the background are removed with the foreground mask. Since our target is to consider both shape and motion feature simultaneously, KLT interest point detector is applied due to its good performance for tracking (Shi and Tomasi, 1994).

After that, small image patches are extracted around the points. An agglomerative clustering algorithm is used to cluster all the patches into several clusters based on a shape descriptor. HOG has been used as the shape descriptor since it is an effective shape descriptor (Dalal and Triggs, 2005). In our evaluations, a cell of 8*8 pixels, a block of 2*2 cells and 9 orientation bins are used for HOG. Hence, the patch size is 16*16 pixels and the final HOG will be a 36 dimensional vector. Euclidean distance is used as the distance measure between two patches.

The spatial occurrence information for each cluster are collected based on a 3*3 blocks as shown in Fig. 1. Each image patch casts a weighted vote for all the clusters based on its location in the training persons. The weights are related to the distance between the patch and the cluster centre. The clusters with small distance will get a higher vote from the image patch.

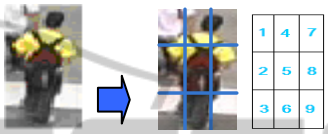


Figure 1: The rectangle is divided into 3*3 blocks, which is used to indicate the patch locations.

Finally, a 3*3 B-ISM (Block-based Implicit Shape Model) is established for a human being. A codebook is formed to save the cluster centres and the spatial occurrence of each cluster. This step is similar as (Hou and Pang, 2010), which gives more details.

3.2 Testing Stage

The details of the method will be introduced in the following three parts: Patch extraction, Shape evidence collection and crowd segmentation.

3.2.1 Patch Extraction

Similar to the training stage, a KLT detector is performed on the testing images. Test patches are extracted around the KLT points in the foreground region. Parameters for HOG descriptor are the same as the training stage.

3.2.2 Evidence Collection

This step would collect spatial information for all the patches in the test image based on the B-ISM established in the training stage.

For each patch, all the codebook entries are searched. The matched entries (the Euclidean distance between the codebook entry and the patch is below a threshold, *th*) will cast a weighted vote based on their similarity. We use $w_{nl}(q_l, c_n) = \exp(-dist(q_l, c_n))$ as the voting

weights, where $dist(q_l, c_n)$ is the Euclidean distance between codebook entry c_n and image patch q_l. Finally, the probability of patch q_l in each block will be obtained with (1). p_{ni} is the probability of the code entry, c_n, in the ith block, which has been saved in the B-ISM. L is the number of patches extracted from the test image. In this way, a 3*3 location table can be obtained for each test patch.

$$p_{li} = \frac{\sum_{dist(q_l,c_n)<th} w_{nl}(q_l, c_n) * p_{ni}}{\sum_{dist(q_l,c_n)<th} w_{nl}(q_l, c_n)}, l = 1,...,L \quad (1)$$

3.2.3 Crowd Segmentation

In our evaluations, a simple rectangle is used as the human model. A set of initial human candidates are nominated based on the points with a sufficiently high occurrence probability in block 1, 4 or 7. A rectangle candidate is proposed with those points as the centre of the top border. Denote the set of nominated rectangles as $R = \{r_k, k = 1,...,K\}$, K is the number of rectangles. The parameters for r_k are the locations and size of the rectangle. Initially, the average human size is used based on its location in the scene. Each initial candidate should have a sufficiently large overlap with the foreground area.

Given a specific configuration, a 2D matrix, $M = \{m_{lk}\}$, is used to indicate the assignments of the KLT points to the candidate rectangles, where $l = 1,...,L$, $k = 1,...,K$. If the interest point l is within the un-occluded region of rectangle k, then $m_{lk} = 1$, otherwise, $m_{lk} = 0$.

Based on its location in the associated rectangles, each KLT point gets a score with (2). p_{li} is the probability of point l in block i, which has been obtained in step-1. $\rho(i,k,l)$ =1 when point l falls inside block i of rectangle k. Otherwise, $\rho(i,k,l)$ =0. The evaluation for the entire crowd configuration, $R = \{r_k, k = 1,...,K\}$, is based on the summation of all the point scores where $s = \sum_{l=1:L} s_l$.

$$s_l = \sum_{k=1:K} (m_{lk} \sum_{i=1:9} (p_{li} \rho(i,k,l))) \quad (2)$$

Starting from the initial set of candidate rectangles, the best configuration is obtained by repeatedly adjusting the candidate size and removing the redundant candidates based on both shape and motion cues. The details are as follows.

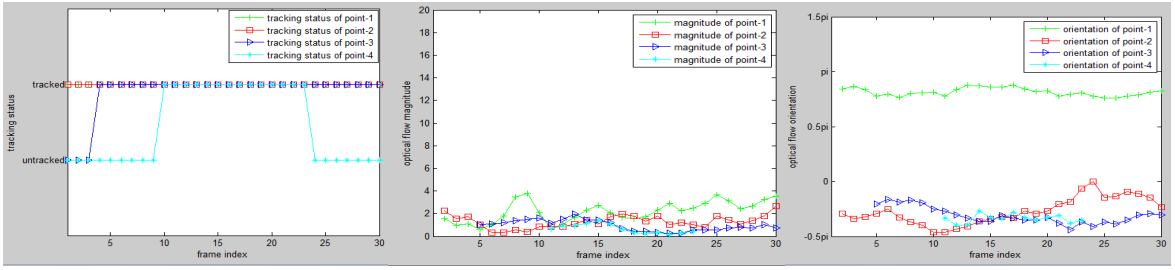
Size Adjustment. For each initial candidate, different scales are tested and the one which can get the highest score for the crowd is used. For simplicity, only the height, h, is adjusted in our evaluations. The best size is picked among 0.8*h, 0.9*h and h.

Removal based on Shape Cues. Two conditions are used for the redundant candidate removal. First, if the entire score can be increased after the candidate removal, then the candidate is removed. Higher score indicates that KLT feature points have been assigned to better locations of the rectangles. Second, candidates with insufficient number of supporting points are removed due to the lack of evidences from the image. Supporting points are defined as those with a decreased score after the candidate removal. In the first loop, any candidate with less than two supporting points is removed. In later stages, a stricter constraint is imposed. A minimal number of supporting points is set for the fully-visible person and candidates have to get enough supporting points to stay on.

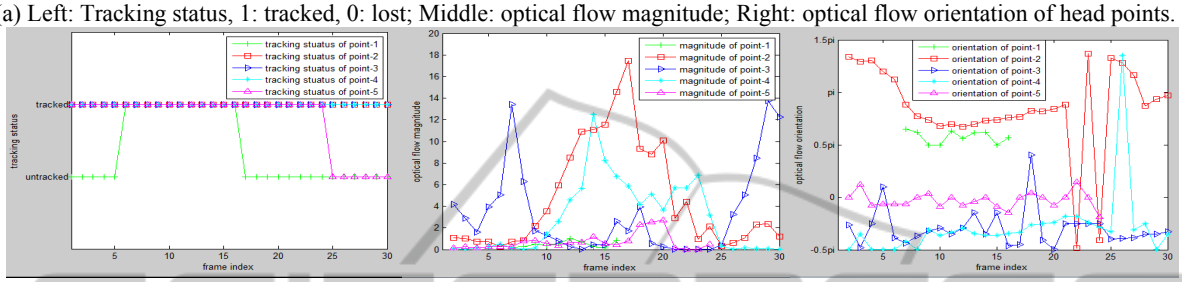
Removal based on Coherent Motion. As shown in (Brostow and Cipolla, 2006, Rabaud and Belongie, 2006), for the general case, points that appear to move together are more likely to be from the same individual. The standard deviation in distance between two KLT points along several consecutive frames can be a measure of the points moving together. Ideally, the distance between two points moving on a rigid object remain the same and the deviation is almost zero.

However, not all the points from the same individual have a low distance deviation. Points on head, torso parts usually show more coherent motion while points on feet or arms often show different trajectories from others. To help the crowd segmentation, a low average standard deviation is expected within each individual and a high average deviation if multiple individuals are considered. Hence, it would be better to use the points with rigid motion only within an individual.

Figure 2: Left: sample points from head, most of which show rigid motion. Right: sample points from feet/arms, most of which show non-rigid motion.



(a) Left: Tracking status, 1: tracked, 0: lost; Middle: optical flow magnitude; Right: optical flow orientation of head points.



(b) Left: Tracking status, 1: tracked, 0: lost; Middle: optical flow magnitude; Right: optical flow orientation of feet/arm points.

Figure 3: Examples of points with rigid motion and non-rigid motion.

Fig. 2 has shown some example points from head and feet/arms. Their tracking status, optical flow magnitude and orientation within +/-15 frames are shown in Fig. 3a and Fig. 3b respectively. Based on the observations, four features are proposed to help distinguish the points with rigid motion: the number of tracked frames within +/-15 frames, the maximal optical flow magnitude, the variation of the optical flow magnitude and the average change of optical flow orientation. Usually, points with rigid motion on a human being can be tracked for a long period. Their optical flow magnitude has small fluctuation and optical flow orientation is almost continuous. For each feature, a threshold is given to define the points with rigid motion. The thresholds are set based on the examination of some sample points. Conservative thresholds are preferred to exclude most points with non-rigid motion, which will lead to a higher credibility of the motion cues. In our evaluations, a point with rigid motion needs to be tracked in more than 24 frames. The maximal optical flow magnitude should be below 5, the variation is below 2.5 and the average optical flow orientation difference is below 0.2.

When only the points with rigid motion are considered, the average deviation should be low for a valid candidate. As we know, the points within a candidate would be assigned to the others after its removal. If the average deviation gets much higher in newly-assigned rectangles, then the candidate must be kept; otherwise, it can be removed in the final results. A margin is set to allow the small fluctuation of the average distance deviation within an individual.

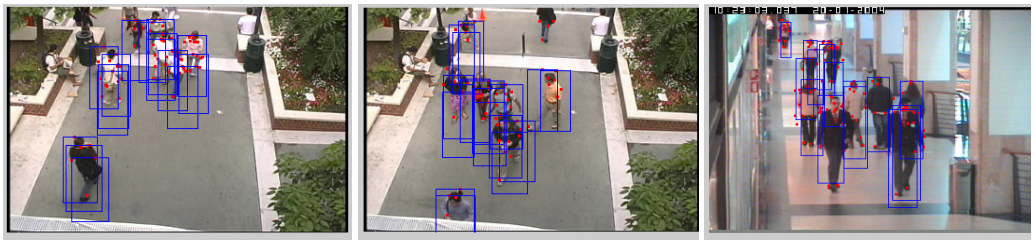
Removal based on Upper-body Rigid Motion. On the human body, most points in the upper body tend to move together, which result in a low average trajectory variation. Hence, a valid candidate should have a low average distance variation in the upper part. A candidate with a large variation is less likely to be a reasonable human person.

With the 3*3 blocks used in Section 3.2, step-1, the average distance deviation of all the points in the top two rows will be examined. The candidate with a very large variation in the upper body will be removed. The threshold is set as a conservative one such that no miss-detection will be caused.

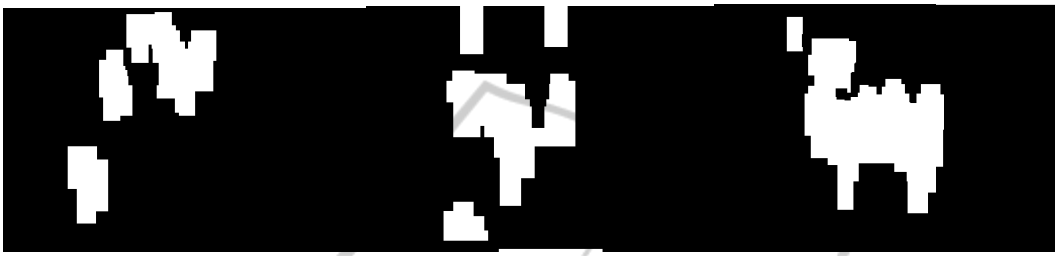
4 EVALUATIONS

'USC-Campus Plaza' and CAVIAR dataset (http://homepages.inf.ed.ac.uk/rbf/CAVIAR/) have been used for the evaluation of the method.

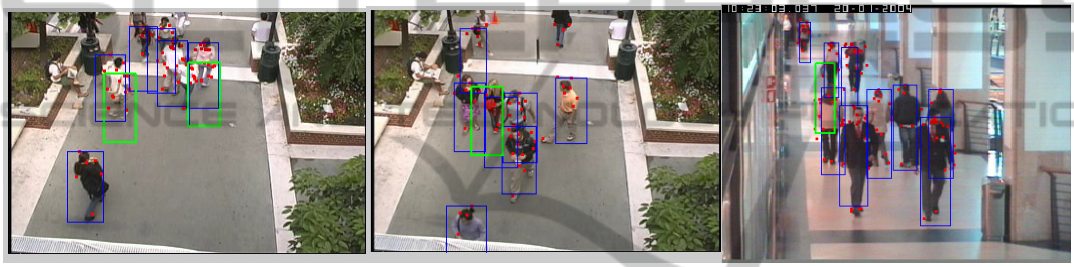
The USC-Campus Plaza sequence was captured from a camera with a 40 degree tilt angle. The frame size is 360*240 pixels and the frame rate is 30fps. It contains 900 frames in total. The training images were extracted from the first 300 frames. 20 training images with 79 persons were used for collecting the training patches. The test images are randomly picked from different periods in the remaining 600 frames and they have different occlusion situations. Most people are different from those in the training set.



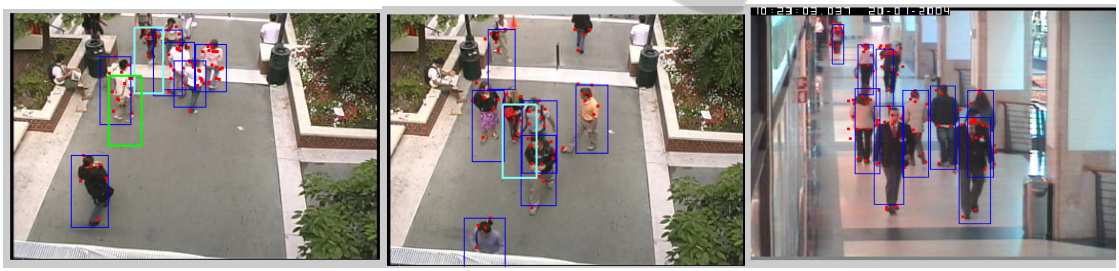
(a) Initial set of candidates. Human is modelled with a rectangle. Red points are the detected KLT points in the foreground.



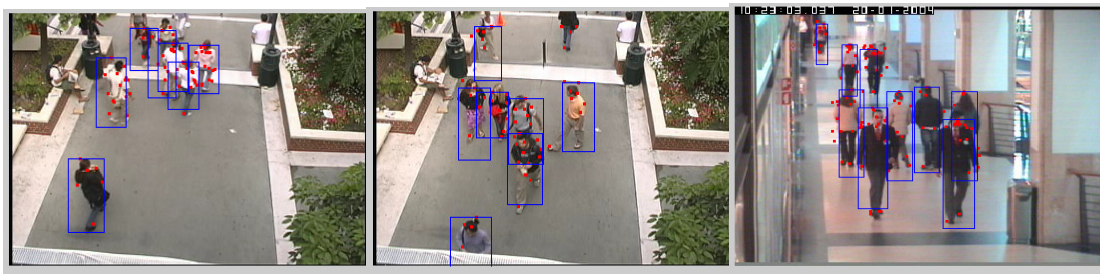
(b) Rough foreground regions used in the evaluations



(c) Results based on shape only. False alarms are indicated in green color, which will be removed using motion features.



(d) Results based on shape and coherent motion. The size of rectangles in cyan colour can be adjusted using upper-body rigid motion.



(e) Results based on shape, coherent motion and upper-body rigid motion.

Figure 4: Selected frames of detection results. Each frame is shown in one column.

Most frames get good enough results based on shape cues only. Three selected sample frames from USC-Campus Plaza and CAVIAR dataset are shown in Fig. 4 to illustrate the improvements after using multi-cues.

Fig. 4a shows the proposed initial candidates. Most persons have got more than one initial rectangle candidates. To show the low requirements

for an accurate foreground contour, a manually obtained rough foreground region is used for each frame in the evaluations, as shown in Fig. 4b. Obviously, it is difficult to get accurate individual detections based on the foreground area only.

The third row is the results based on shape cues only. Most individuals can be located well based on shape cues. However, cues based on shape may be less reliable when the crowd is dense, the background is complicated or other human shape-like region appears. Hence, some false detections may stay in the shape-based results, which are indicated in green colour in Fig. 4c. The fourth row is the results when motion consistency is also considered. It can be seen that the false detections indicated with the green colour in the third row have been removed based on the coherent motion rule in Fig. 4d. In the left column, three close persons have got better detection results based on their different trajectories. Similarly, in the middle column, a false candidate covering two close persons has been removed. Finally, the bottom row shows the results when the upper-body rigid motion is also considered. In the left column of Fig. 4e, a false candidate with high trajectory variations in the upper part has been removed. In addition, better human size has been obtained for the persons indicated in cyan colour in Fig. 4d.

5 CONCLUSIONS

In this paper, a method based on both shape and motion features for crowd segmentation is presented. The shape-based method has formulated the problem into a feature point clustering process. Multi-frame coherent motion of the feature points on a person is used to enhance the segmentation performance. Most feature points on the human upper-body are moving together, which are used to get more reasonable detections.

REFERENCES

Brostow, G. J. & Cipolla, R. 2006. Unsupervised Bayesian Detection of Independent Motion in Crowds. IEEE Conference on Computer Vision and Pattern Recognition.

Dalal, N. & Triggs, B. 2005. Histograms of oriented gradients for human detection. IEEE Conference on Computer Vision and Pattern Recognition.

Dalal, N., Triggs, B. & Schmid, C. 2006 Human detection using oriented histograms of flow and appearance. European Conference on Computer Vision.

Hou, Y.-L. & Pang, G. K. H. 2009. Human Detection in a Challenging Situation. IEEE International Conference on Image Processing.

Hou, Y.-L. & Pang, G. K. H. 2010. Human Detection in Crowded Scenes. IEEE International Conference on Image Processing.

http://homepages.inf.ed.ac.uk/rbf/caviar/.

Lin, Z., Davis, L. S., Doermann, D. & Dementhon, D. 2007. Hierarchical Part-Template Matching for Human Detection and Segmentation. IEEE International Conference on Computer Vision.

Rabaud, V. & Belongie, S. 2006. Counting Crowded Moving Objects. IEEE Conference on Computer Vision and Pattern Recognition.

Rittscher, J., TU, P. H. & Krahnstoever, N. 2005. Simultaneous estimation of segmentation and shape. IEEE Conference on Computer Vision and Pattern Recognition.

Shi, J. & Tomasi, C. 1994. Good features to track. IEEE Conference on Computer Vision and Pattern Recognition.

Tuzel, O., Porikli, F. & Meer, P. 2007. Human Detection via Classification on Riemannian Manifolds. IEEE Conference on Computer Vision and Pattern Recognition.

Viola, P., Jones, M. J. & Snow, D. 2003. Detecting pedestrians using patterns of motion and appearance. IEEE International Conference on Computer Vision.

Wang, X., Han, T. X. & Yan, S. 2009. An HOG-LBP Human Detector with Partial Occlusion Handling. IEEE International Conference on Computer Vision.

Zhao, T. & Nevatia, R. 2004. Tracking multiple humans in complex situations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26, 1208-1221.

Zhao, T., Nevatia, R. & Wu, B. 2008. Segmentation and Tracking of Multiple Humans in Crowded Environments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30, 1198-1211.

Zhu, Q., Yeh, M.-C., Cheng, K.-T. & Avidan, S. 2006. Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. IEEE Conference on Computer Vision and Pattern Recognition.