



Title	Multicue-based crowd segmentation using appearance and motion
Author(s)	Hou, YL; Pang, GKH
Citation	IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems & Humans, 2013, v. 43 n. 2, p. 356-369
Issued Date	2013
URL	http://hdl.handle.net/10722/164089
Rights	IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems & Humans. Copyright © IEEE.

Multicue-Based Crowd Segmentation Using Appearance and Motion

Ya-Li Hou and Grantham K. H. Pang, *Senior Member, IEEE*

Abstract—In this paper, our aim is to segment a foreground region into individual persons in crowded scenes. We will focus on the combination of multiple clues for crowd segmentation. To ensure a wide range of applications, few assumptions are needed on the scenarios. In the developed method, crowd segmentation is formulated as a process to group the feature points with a human model. It is assumed that a foreground region has been detected and that an informative foreground contour is not required. The approach adopts a block-based implicit shape model (B-ISM) to collect some typical patches from a human being and assess the possibility of their occurrence in each part of a body. The combination of appearance cues with coherent motion of the feature points in each individual is considered. Some results based on the USC-Campus sequence and the CAVIAR data set have been shown. The contributions of this paper are threefold. First, a new B-ISM model is developed, and it is combined with joint occlusion analysis for crowd segmentation. The requirement for an accurate foreground contour is reduced. In addition, ambiguity in a dense area can be handled by collecting the evidences inside the crowd region based on the B-ISM. Furthermore, motion cues—which are coherent moving trajectories of feature points from individuals—are combined with appearance cues to help segment the foreground region into individuals. The usage of motion cues can be an effective supplement to appearance cues, particularly when the background is cluttered or the crowd is dense. Third, three features have been proposed to distinguish points on rigid body parts from those with articulated movements. Coherent motion of feature points on each individual can be more reliably identified by excluding points with articulated motion.

Index Terms—Crowd segmentation, implicit shape model (ISM), independent motion, occlusions.

I. INTRODUCTION

PEOPLE COUNTING and human detection are two important problems in visual surveillance. It is useful for shopping mall managers to have information on the number of people in a mall each day. In addition, to ensure the safety of people and facilities, video surveillance has become more and more important. In a video surveillance system, detecting

individuals is usually the first step for further analysis. After finding human beings in the image, posture estimation and body part segmentation can be performed for better understanding of the scenario [1]–[4].

Much interesting work on human detection has been carried out in recent years. People counting and individual detection have been achieved in two ways. The first method is to exhaustively search human beings with a sliding window. Different features have been explored for classification of human and nonhuman. However, the method is usually computationally expensive. In the other method, individual detection is achieved by crowd segmentation. The computation load is relatively lower. However, most segmentation-based methods rely on an informative foreground contour. The goal of this paper is to develop a segmentation-based method based on both appearance and motion cues for individual detection.

II. RELATED WORK

People counting and human detection have become hot topics in recent years. Currently, a great deal of research has been carried out in these areas. The methods based on appearance cues can be classified into two main categories.

The first category usually searches an image exhaustively with a sliding window. Each window is classified as human or nonhuman based on the features of shape, color, or texture [5]–[8]. These methods are usually extended by considering the motion feature between two consecutive frames [9], [10]. To reduce the number of scanned windows, Li *et al.* [11] searched the head-shoulder shape, based on the Histogram of Oriented Gradients (HOG) descriptor, inside the foreground region. However, the methods in this category are still computationally expensive and usually require a high-resolution image.

The second category assumes that a foreground area for the crowd has been obtained. Human detection is then simplified as a problem to segment the foreground into individuals [12]–[17]. Our method belongs to the second category, and the related work will be reviewed in the following paragraphs.

Zhao and Nevatia [12]–[14] locate the individuals in the foreground area by head detection. In their early work [13], head candidates were detected by checking local peaks on the foreground contour. A detected individual was removed from the foreground, and the rest of the individuals were detected in the remaining foreground region. In their later work [14], a simple “ Ω ” template was also considered for head detection inside the foreground area. A more sophisticated sampling algorithm, i.e., Markov chain Monte Carlo, was used to find the optimal crowd configuration.

Manuscript received September 28, 2011; revised February 19, 2012; accepted April 1, 2012. Date of publication September 12, 2012; date of current version February 12, 2013. This work was supported in part by The University of Hong Kong under Committee on Conference and Research Grant 21476021 and in part by Beijing Jiaotong University through the Startup Funding. This paper was recommended by Associate Editor J. Wu.

Y.-L. Hou is with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China (e-mail: ylhou@bjtu.edu.cn).

G. K. H. Pang is with the Industrial Automation Research Laboratory, Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam, Hong Kong (e-mail: gpang@eee.hku.hk).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TSMCA.2012.2199308

Rittscher *et al.* [15] tried to reduce the requirements for an accurate foreground contour by sampling only the informative feature points from the contour. The sampled points were labeled as top, bottom, left, and right, which was based on their local contour information. A variant of expectation (E)–maximization (M) algorithm has been used to find the best grouping of the points within rectangles. In the E step, feature points are assigned to the rectangle candidates with a probability based on the distance to the corresponding top, bottom, left, or right borders of the rectangles. In the M step, rectangle sizes and locations are updated based on their associations with the feature points. Points with a low assignment probability have a low influence on the rectangles.

The work of Dong *et al.* [17] is also based on the insight that the foreground contour is a strong indication about the number and positions of human beings in a crowd. Since the number of people inside a foreground blob will never suddenly change, ambiguities inside the crowd region are mitigated by considering a set of consecutive frames. However, this strategy does not solve the problem inherently; for example, when the crowd does not show significant movement, ambiguities inside a dense crowd will cause a significant drop in the performance of the method. Another fatal problem of the method is that a great number of labeled training samples are required. To include almost all the possible occlusion situations, a small number of people will need a huge number of training samples. The requirements will seriously limit the application of the method in a large crowd.

In summary, the methods described previously have three serious limitations: First, they usually rely on an informative foreground contour, which cannot be easily obtained in most situations. An example scenario is in [16], where most people are moving slightly and showing only scattered foreground pixels. Second, with little information inside the foreground area, the methods would find it difficult to handle the high ambiguity at the center of the dense crowds, which limits their applications in dense scenarios. Third, the methods assume that the region shown in the foreground is from human beings. No specific measure has been taken to deal with nonhuman objects in the foreground. In addition, except for the background subtraction, the methods in [12]–[15] do not consider the motion features explicitly. Although the temporal information has been used in [17] to handle the ambiguities inside the crowd region, it has not solved the problem inherently. If no significant information can be extracted about the number of people on the foreground contour for the entire sequence, people in the dense region will never be counted.

In [16], individual detections are carried out by clustering cornerlike feature points within the foreground area. By using the feature points from the foreground objects, reliance on an accurate foreground contour is reduced. Some good individual detection results have been achieved in challenging situations. However, information for clustering points in a dense area is still not sufficient.

The implicit shape model (ISM) in [18] and [19] has been shown to be an effective model based on the local appearances of human beings. With the ISM, more evidence can be collected for the crowd configuration in a dense area. This model

specifies where the local appearances might occur with respect to the object center. This was used initially for car and cow detection in [18]. In [19], it was used for pedestrian detection. Some small patches are extracted around the interest points, and they are used to vote for the human centers. The maxima in the 3-D voting space are searched with the mean-shift algorithm to form an initial hypothesis. At the verification stage, chamfer matching and a minimum-description-length-based analysis are performed to refine the initial hypothesis. However, the method might not work well in a crowded situation. When a person is seriously occluded, he or she might not be able to get enough votes to appear in the initial hypothesis. Only results with slight occlusions are shown in [19].

In addition to the study on human shape, some impressive motion characteristics have been observed for human detection in Brostow and Cipolla [20] and Rabaud and Belongie [21]. In both papers, it is believed that points from the same person display consistent trajectories, while points from different persons usually have different moving 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 pedestrians exhibit sustained articulations. In addition, very little appearance information has been explored in these methods. As far as we know, there has been little work on combining the multiframe motion features with an appearance-based method.

Motivated by the aforementioned work, individual detection is formulated as a problem of feature point clustering in this paper. A modified ISM of a human being, called the block-based ISM (B-ISM), is established to provide sufficient evidence for the crowd configuration. Coherent motion of the points from the same person will be used as a supplement to the appearance-based method. Our aim is to propose a method to use multiple clues simultaneously for crowd segmentation, which have few constraints on the scenarios to ensure a wide range of applications. Details are introduced in Section III.

III. METHOD

The developed approach includes two stages, as shown in Fig. 1. In the training stage, a codebook consisting of some typical local human appearances is formed. A B-ISM is established by collecting information on where each codebook entry might occur on the human body. In the testing stage, individual segmentation is performed based on both appearance and motion cues. The established B-ISM will be used as appearance cues to group the extracted patches in a test image. Motion cues can be obtained from coherent moving trajectories of feature points from individuals. These cues are then combined with appearance cues to help segment the foreground region into individuals. The usage of motion cues can be an effective supplement to the appearance cues, particularly when the background is cluttered or the crowd is dense.

A. Training Stage

Human appearances might look different in 2-D images when the camera viewing angles are different. Hence, a training

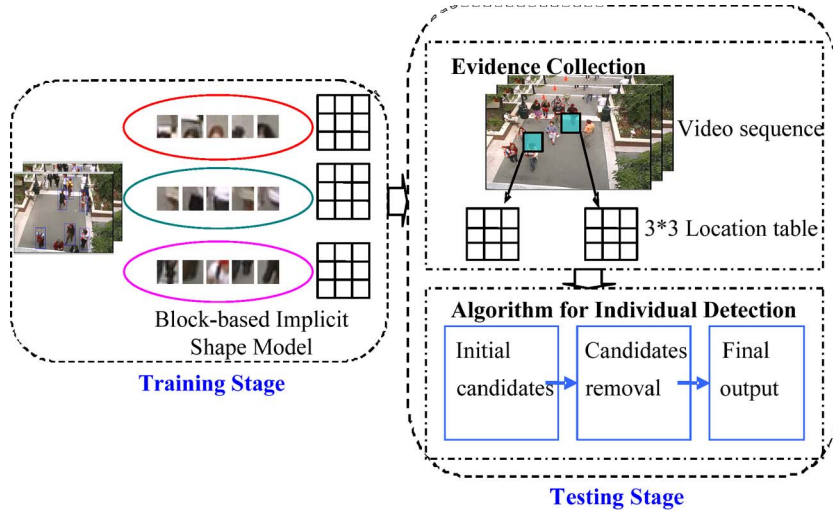


Fig. 1. Structure of the developed method.



Fig. 2. Examples of training images. (a) Training persons are annotated with a rectangular bounding box. (b) Foreground region for training persons.

process is necessary for a specific camera setup to learn the knowledge of human appearances in the 2-D space.

The training images contain some fully visible human beings. For those selected training persons, a bounding box is annotated, indicating human locations and sizes. It is also assumed that we have the foreground region of the training persons. An example of training images used in the USC-Campus sequence is shown in Fig. 2. Manual annotation is carried out only once for a fixed camera setup.

The training stage has three steps: image patch extraction, patch clustering, and formation of B-ISM. The average human size at different locations in the scene is also estimated based on the training persons.

Step 1—Patch Extraction for B-ISM: In this step, interest points are detected from the training persons. Small image patches are extracted with each interest point as the center. In [19] and [22], different combinations of four interest point detectors and two local appearance descriptors were evaluated. The results showed that the Hessian–Laplace detector and shape context descriptors have the best performance for pedestrian detection.

In our method, the Kanade–Lucas–Tomasi (KLT) [23], [24] interest point detector is employed. This is because the KLT detector can provide a good feature for tracking. Briefly, the features are detected by examining the minimum eigenvalue of the Hessian matrix within a small window at each location. The window size is set at 7×7 pixels in our evaluations. A shape-context-like descriptor, i.e., HOG, has been used as the appearance descriptor of each image patch. As suggested in [5],



Fig. 3. Example of KLT detection and patch extraction.

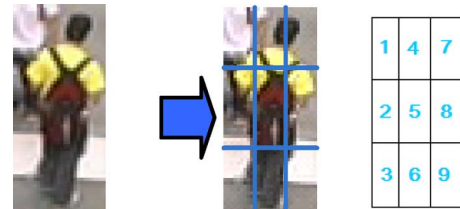


Fig. 4. Patch location is indicated with the 3×3 blocks in a rectangle.

a cell of 8×8 pixels and a block of 2×2 cells were used. The gradient orientation is divided into nine orientation bins. Each image patch is 16×16 pixels, and the dimension of the HOG vector is 36.

A number of patches can be collected based on all the training images. Fig. 3 shows an example of a training person, the detected KLT points on this person, and the extracted patches. It can be observed that some KLT points are from human contours while some of them are due to clothes, backpacks, and so on.

As mentioned, each selected training person is shown in a rectangular shape that can be divided into a 3×3 block. The index of each block is shown in Fig. 4. Hence, each extracted patch has an associated location index, which depends on the block where it belongs in the 3×3 block of the training person. All the extracted patches can be denoted with a set of pairs $\{(q_l, b_l), l = 1, 2, \dots, L\}$. q_l is the HOG vector of each patch, and b_l is its associated block index.

Step 2—Patch Clustering: The aim of this step is to group the extracted patches in Step 1 into several clusters based on their appearances. An agglomerative clustering algorithm in [18] is used in our evaluations.

Starting with all the extracted patches as a separate cluster, the clustering process continues by merging the two most

similar clusters together at each step. The Euclidean distance between the appearance descriptor vectors is used as the similarity measure of two image patches, denoted as $dist(q_1, q_2)$. The similarity distance between two clusters is calculated by (1), in which cluster C_1 contains $|C_1|$ patches and C_2 has $|C_2|$ patches. The merging process ends when the distance between the two most similar clusters is above a threshold $th = 0.7$. Clusters with only very few samples are removed.

$$dist(C_1, C_2) = \frac{\sum_{q_1 \in C_1, q_2 \in C_2} dist(q_1, q_2)}{|C_1| * |C_2|}. \quad (1)$$

A good clustering result is important for the ISM-based method. Hence, we suggest a user-guided clustering method to improve the clustering result. Although this will be a semiautomatic method, it is usually acceptable since the training process is performed only once. The automatic clustering results from the agglomerative algorithm are further checked. A cluster that contains very different types of patches is further divided into multiple new clusters. Several representative patches from the cluster are used as the initial seeds of these new clusters. The remaining patches are grouped to the closest cluster.

In the evaluation of USC-Campus sequence, 648 patches were extracted based on 76 training persons from 20 training images. Thirteen clusters were obtained based on the agglomerative algorithm. After a manual inspection of the clusters, 21 clusters were obtained eventually.

Step 3—B-ISM: The average of the 36-D HOG features of all the patches in each cluster is calculated as the center of the cluster. The cluster centers are then stored as codebook entries in B-ISM to represent the cluster. Suppose that a total of N clusters have been obtained; to establish a B-ISM, we also have to collect the spatial occurrence information of the codebook entries on the human body.

The spatial occurrence information of each cluster is collected with a voting stage. For each extracted patch q_l from Step 1, its location has to be registered with the activated entry. In our evaluations, each patch casts a vote to the closest cluster with a weight of one. The total weights in each block for each cluster can be calculated with (2). In (2), $\delta(q_l, c_n) = 1$ only when n is the cluster with the minimum distance to q_l ; otherwise, $\delta(q_l, c_n) = 0$.

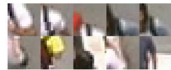


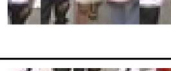
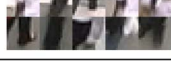
$$s_{ni} = \sum_{l=1 \dots L, b_l=i} \delta(q_l, c_n), \quad n = 1, \dots, N; \quad i = 1, \dots, 9. \quad (2)$$

After the voting stage, nine spatial occurrence values are obtained for each codebook entry, corresponding to the nine block locations. Finally, the nine values for each entry are normalized to get the probability of the entry in each block

$$p_{ni} = \frac{s_{ni}}{\sum_{i=1, \dots, 9} s_{ni}}, \quad n = 1, \dots, N; \quad i = 1, \dots, 9. \quad (3)$$

At the end of the training stage, a human model with N typical local appearances and their spatial occurrence probabilities in each block is established. To distinguish the new model from the one in [19], we called it a B-ISM.

TABLE I
SAMPLE CLUSTERS AND THEIR OCCURRENCE TABLE

Index	# patch in the cluster	Sample patches	3x3 spatial occurrence table
1	29		0.0256 0.1282 0.3846 0.0256 0.1538 0.2051 0 0 0.0769
2	16		0.4737 0.0526 0.0526 0.2105 0.0526 0 0.0526 0.1053 0
3	35		0 0.6364 0.0606 0 0.0606 0.1515 0.0303 0.0606 0
4	11		0.0667 0.0667 0 0 0.6667 0.1333 0 0 0.0667
5	38		0.2703 0.0541 0 0.1351 0.0541 0.1892 0 0.2703 0.0270

Some sample clusters obtained from the USC-Campus sequence are shown in Table I. The first cluster has shoulderlike shapes, and as expected, the spatial occurrence table has a higher probability in block 7, which is 0.3864. For the third cluster, most patches are from head top, and the spatial occurrence table indicates a higher probability in block 4. Due to the special viewing angle, most legs tend to be left tilted. As shown in the fifth cluster, edges with a left tilted angle could be from left shoulder or leg parts, which have an equal probability of 0.2703.

Step 4—Average Human Size: In the USC-Campus sequence, human scales are related to both coordinates of the image. In the evaluations, 189 training persons from different locations are used to learn the average human size in the scene.

The entire image is divided into 12×18 blocks, and each block has 20×20 pixels. First, an average human size in each block is calculated based on all the training persons. Then, linear interpolation is repeatedly performed along each row and column a couple of times. This will help get size estimation for those blocks without training persons. At the same time, alias of sizes in blocks with training persons can be reduced.

Finally, the average human scales in each block are stored in a table for use in later stages. In the USC-Campus scene, people at the bottom left corner are the largest, and the people far away tend to be thinner.

B. Testing Stage

The testing stage aims to segment the crowd region into individuals based on both appearance and motion cues. It is assumed that a crowd region has been obtained. It must be emphasized that it is easier to obtain a foreground region than an accurate foreground contour. For example, a foreground region for the slightly moving people is obtained by a closing operation on the scattered foreground pixels [25]. Another method to detect crowd regions in still images is described in [26].

After getting a foreground region, crowd segmentation is implemented with the following steps: patch extraction, evidence collection, and individual segmentation.

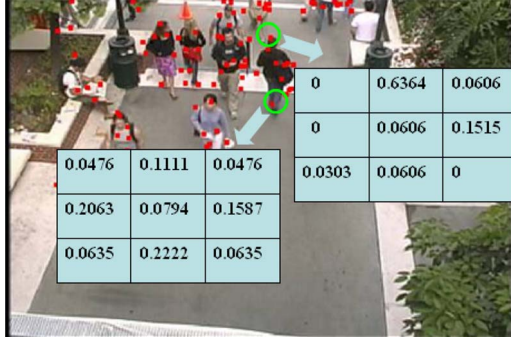


Fig. 5. Each KLT point has obtained a 3×3 spatial occurrence table. Usually, the head point has a higher probability in the top row, while the feet point has a higher probability in the bottom row.

Step 1—Patch Extraction: Operations similar to those of Step 1 in the training stage are performed on the test image. A KLT interest point detector is applied, and small image patches are extracted around the KLT points. The same parameter setup should be used as the one in the training stage. Suppose that a total number of L image patches have been extracted in the foreground area.

Step 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. Given a patch q_l ($l = 1, \dots, L$, and L is the number of patches extracted from the test image), all the codebook entries are searched. The occurrence location table for the patch is collected based on the activated entries. In our evaluations, only the closest cluster is activated and casts a vote to the patch with a weight of one. Hence, the probability of patch q_l to occur in each block is obtained by (4). In (4), $\delta(q_l, c_n) = 1$ only when c_n is the cluster with the minimum distance to q_l ; otherwise, $\delta(q_l, c_n) = 0$

$$p_{li} = \sum_{n=1 \dots N} p_{ni} \delta(q_l, c_n), \quad i=1, \dots, 9; \quad l=1, \dots, L. \quad (4)$$

In this way, a 3×3 location table can be obtained for each test patch. This table indicates the probability that a patch might come from when considering the nine block locations. Usually, head points will get a higher probability value in the top row (block locations 1, 4, and 7), while points from the feet region will get a higher probability in the bottom row (block locations 3, 6, and 9). Fig. 5 has shown an example of the 3×3 location table for head and feet points.

Step 3—Crowd Segmentation:

Formation of initial rectangle candidates: In most situations, the human's head region is visible when he is in the scene. Hence, a set of initial human candidates will be nominated based on the head regions in our evaluations. Based on the occurrence location tables from Step 2, head regions can be detected from the points with a high probability in block 1, 4, or 7. In the method, a rectangle is used as the human model. When a patch has a probability above 0.112 ($\approx 1/9$) to occur in these locations, an initial rectangle candidate is proposed. The rectangle is set with the head point as the center of the top border. Fig. 6(a) shows two examples of forming rectangle

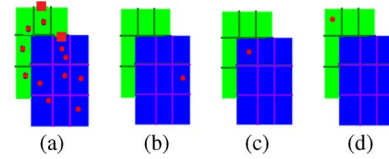


Fig. 6. (a) Examples of forming rectangle candidates by the head region points. (b)–(d) Each KLT feature point is assigned to a rectangle based on the occlusion map, and a score is obtained based on its location in the rectangle. Denote the blue rectangle with $k = 1$ and the green rectangle with $k = 2$; then, $\rho(8, 1, l) = 1$ in (b), $\rho(1, 1, l) = 1$ in (c), and $\rho(1, 2, l) = 1$ in (d).

candidates by the head region points (indicated with squares). 0.112 is a very conservative threshold to ensure that all the good candidates are included in the initial set of rectangles.

Rectangle size is initially set as the average human size based on head location. The set of nominated rectangles is denoted as $R = \{r_k, k = 1, \dots, K\}$, and K is the number of rectangles. The parameters for r_k are the locations and size of rectangle k . To reduce the number of initial candidates, some measures have been taken to restrict the nomination process. First, each candidate should have sufficiently large overlap with the foreground area. Second, a minimal distance among the head point candidates is assured. Third, the central area of a valid candidate should be a foreground area.

Assignment of KLT points to the rectangle candidates: Given a specific configuration, a 2-D matrix $M = \{m_{lk}\}$ is used to indicate the assignments of the KLT points to the rectangle candidates, where $l = 1, \dots, L$ and $k = 1, \dots, K$. As shown in Fig. 6(a), it is reasonable to assume that rectangles with lower y -coordinates in an image are occluded by those with higher y -coordinates in the overlapped region. If the interest point l is within the nonoccluded region of rectangle k , then $m_{lk} = 1$; otherwise, $m_{lk} = 0$. To allow for a small deviation of the KLT points, point l outside the candidate k but within a small margin also has $m_{lk} = 1$. Finally, M is normalized along each row.

Calculation of score for each KLT point: Based on the location of a KLT point in each of the assigned rectangles, a score is calculated for the point by (5). Since the occurrence location table from Step 2 indicates the probability that the patch might occur in different human parts, the score is a measure of accuracy on each KLT point falling into the location that is desirable. It is higher when the patch falls in a reasonable location in the associated rectangle.

$$s_l = \sum_{k=1:K} \left(m_{lk} \sum_{i=1:9} (p_{li} \rho(i, k, l)) \right). \quad (5)$$

In (5), $m_{lk} > 0$ only when point l is assigned to rectangle k . p_{li} is the probability of point l to occur in block i , which has been obtained in Step 2. The index $\rho(i, k, l) = 1$ only when point l falls in block i of rectangle k or within a small margin; otherwise, $\rho(i, k, l) = 0$. Fig. 6(b)–(d) shows some illustrations of this issue. Denote the blue rectangle with $k = 1$ and the green one with $k = 2$. $\rho(8, 1, l) = 1$ in Fig. 6(b), $\rho(1, 1, l) = 1$ in Fig. 6(c), and $\rho(1, 2, l) = 1$ in Fig. 6(d).

The evaluation of the entire crowd configuration $R = \{r_k, k = 1, \dots, K\}$ is based on the summation of all the point

scores, as shown in

$$s = \sum_{l=1:L} s_l. \quad (6)$$

Removal of redundant rectangle candidates: Candidates in the initial set are examined in the descending order of their y -coordinates. The best configuration is obtained by the following iterative steps. In each step, the configuration is updated by repeatedly adjusting the candidate size and removing the redundant ones.

- 1) **Size adjustment.** For each initial candidate, different scales are tested, and the one that produces the highest score in (6) is used. For simplicity, only the height h is adjusted in our evaluations. The best size is picked among $0.8 \times h$, $0.9 \times h$, and h for the USC-Campus sequence and $0.9 \times h$ and h for the CAVIAR set.
- 2) **Foreground cues.** Foreground area is a strong evidence of a candidate. Assuming that candidates with a lower y -coordinate are occluded by those with higher y -coordinates, foreground region is assigned to each rectangle. After the removal of a candidate, some foreground area might fall out of all the rectangles, which are called support area of the candidate. When the support area is sufficiently large, it is better to keep the candidate. In the USC-Campus sequence, when the support area is larger than 45% of the candidate, the candidate is retained.
- 3) **Appearance cues.** In a crowd configuration, KLT points are assigned to each rectangle based on the occlusion map. After the removal of a candidate, the points inside the candidate will be assigned to new rectangles, and new scores can be obtained by (5). The points with a decreased score after the removal of a candidate are defined as support points for the candidate. In other words, the support points can get a more reasonable block location in the original candidate based on the B-ISM model.

In our evaluations, two conditions are used for the redundant candidate removal. First, if the total score in (6) is increased after the candidate removal, then the candidate is removed. Second, candidates with an insufficient number of support points will be removed due to the lack of evidence from the image. To increase the reliability of the support points, the points very close to other persons are not counted.

The candidate examination based on appearance cues is performed in two stages. In the first stage, any candidate with less than two support points is removed. The most trivial candidates can be removed in this stage. In the second stage, a further analysis is examined. The minimum number of support points for a fully visible person is defined as N_s^2 . The minimum number of support points for a partially visible persons is $(percentage\ of\ visible\ area) \times N_s^2$. Feature points on a human body might come from the human contour, clothes, and other decorations. Since clothes might show a different number of feature points, feature points from human contour are most indicative. N_s^2 is mainly defined based on the number of feature points from human con-



Fig. 7. (a) After the removal of the green rectangle, points in the green candidate originally will be assigned to the blue rectangle, in which points from two persons are grouped together. (b) After the removal of the green rectangle, points in the green candidate originally will be assigned to the two blue rectangles, in which points are all from one individual.



Fig. 8. Sample points from (a) head/torso and (b) feet/arms.

tour, which is related to image resolution, camera viewing angle, and distance of objects from the camera. In the evaluation of the USC-Campus sequence, $N_s^2 = 4$, which is a conservative value to avoid missed detections.

- 4) **Motion cues.** As shown in [20] and [21], points that appear to move together are more likely to be from the same individual. The standard deviation in the 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 remains the same, and the deviation is almost zero.

Due to coherent motion of feature points in an individual, the distance deviation among the KLT points in a valid candidate should be low. When a candidate is removed, the points within the candidate would be assigned to other candidates. If motion coherence in newly assigned rectangles is getting much worse, the points in the new rectangle are likely to be from a different person. Hence, it is better to retain the original candidate. Fig. 7(a) shows an example of this case. On the other hand, if the moving trajectories in the newly assigned rectangles are even more consistent, it is better to group the points with the new rectangle. Hence, the original candidate should be removed. Fig. 7(b) shows an example of this case.

However, it should be noted that points with articulated motion, particularly from feet and arms, often show different trajectories. As a result, even the distance deviation among points inside one candidate can be large. For crowd segmentation, a low average standard deviation is expected within each individual, and a high average deviation is expected for multiple individuals. Hence, it would be better to use points from a rigid body part only for the calculation of coherent motion.

Points in Fig. 8(a) are sample points from heads and torsos, which move more like a rigid body. Points in Fig. 8(b) are from feet or arms, which usually have some

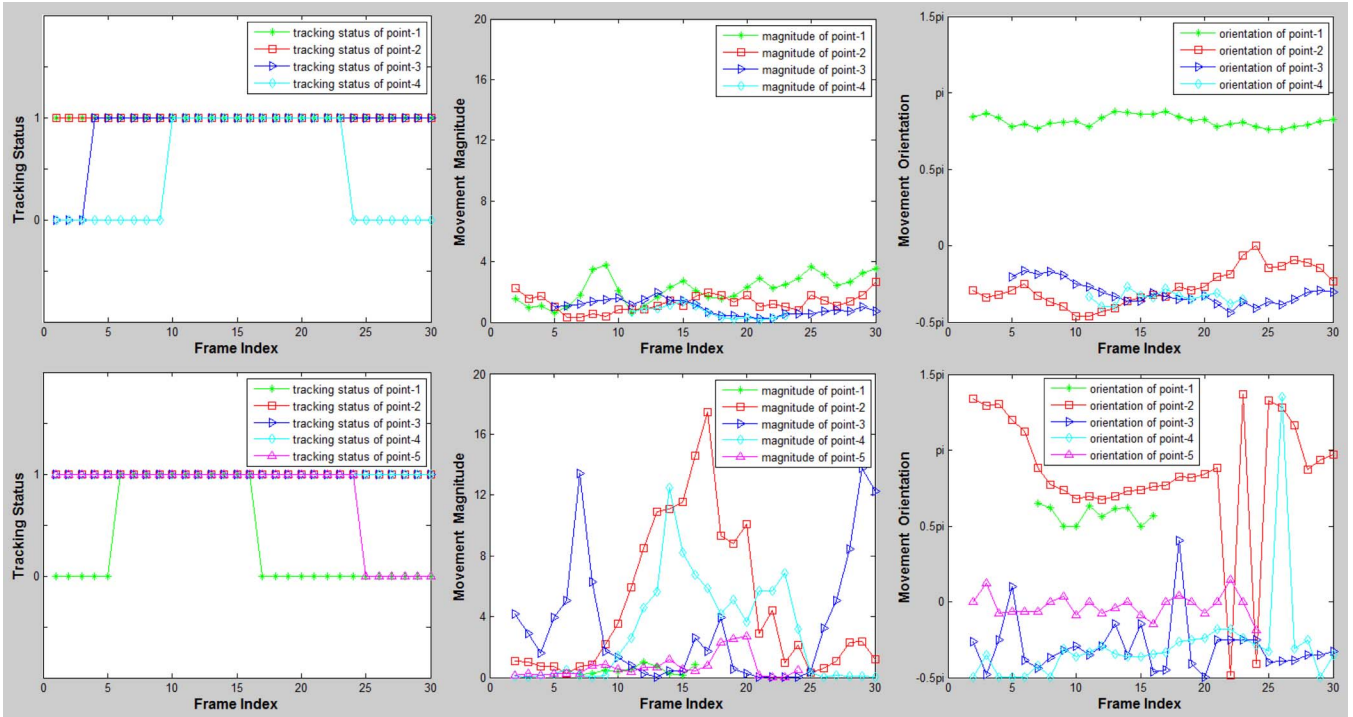


Fig. 9. Movement of sample points from head/torso and feet/arm parts. (Top) Tracking status, movement magnitude, and orientation of head/torso points. (Bottom) Tracking status, movement magnitude, and orientation of feet/arm points. For tracking status, “1” means tracked, and “0” means lost.

articulated movements. Their tracking status, movement magnitude, and orientation along the previous and next 15 frames are shown in Fig. 9. Tracking status is indicated with “1” when the point is successfully tracked and “0” if it is lost. Due to less local deformation, most points from head and torso can be tracked for a long period. Their movement magnitude between two consecutive frames has small fluctuation, and the movement orientation is almost continuous. On the other hand, due to frequent local deformation, points with articulated motion might be lost easily. Even if they can be tracked, movement magnitude variation is large, and there are often some sudden changes of movement orientation.

Based on these observations, three features are proposed to help distinguish the points from the rigid body parts:

- tracking status: the number of frames tracked successfully within the previous and next 15 frames n_f ;
- movement magnitude: standard deviation of movement magnitude during the tracked period σ_m ;
- movement orientation: average change of movement orientation during the tracked period \overline{m}_{do} .

For each feature, a threshold is given to define the points moving on a rigid part. The thresholds are related to the frame rate, the viewing angle, and the distance of objects from the camera. In our evaluations, the thresholds are set based on an examination of some sample points. Conservative thresholds are preferred so as to retain the points that are more likely to be on a rigid body part. In the evaluations of the USC-Campus sequence, a point moving on a rigid part needs to be tracked for more than 24 frames, σ_m is below 2.5, and \overline{m}_{do} is below 0.2.

- Candidate removal.** Finally, based on different confidence levels, candidates are examined in the following steps. First, if the support area of a candidate is larger than 45%, the candidate must be retained. Second, if the score of the entire configuration is decreased after the removal or the support points are equal to or less than one, then it is better to remove the candidate. Third, if the trajectory deviation gets much higher in newly assigned rectangles, then the candidate is retained. Fourth, if the number of support points is not sufficient or the trajectory deviation is similar or even lower, then the candidate is removed.

The details of the implementation of the algorithm are shown in Table II.

IV. EVALUATIONS

An implementation of the KLT interest point detector in [27] was used in our evaluations. The evaluations include three parts:

- test of the appearance-based method using the USC-Campus sequence;
- comparison of appearance- and multicue-based methods based on the same USC-Campus sequence;
- test of both appearance- and multicue-based methods based on the CAVIAR video sequences.

A. Crowd Segmentation Based on Appearance Cues

The USC-Campus video sequence was an outdoor scene on a campus. It was captured from a camera with a 40° tilt angle. The frame size is 360×240 pixels, and the frame rate is 30 fps. It contains 900 frames in total.

TABLE II
ALGORITHM FOR CROWD SEGMENTATION

Initialization:

Initial rectangles are nominated. All the rectangles are sorted in descending order according to their y-coordinates. $R = \{r_k, k = 1, \dots, K\}$.

Each point gets a score $S = \{s_1, s_2, \dots, s_l\}$ and the total score $s = \sum_{l=1}^L s_l$.

Stage-1: Loop until the rectangles are not changed.

Repeat for $k=1 \dots K$

(a) *Size adjustment.* Several scales are tried and the best one which gets the highest score is adopted. If the new total score $s' > s$, then the rectangle set and the scores are updated. $R \leftarrow R'$, $s = s'$, $S \leftarrow S'$.

(b) *Appearance & motion cues.* In $R' = R - r_k$, based on the occlusion map,

- Reassign the foreground area to the rectangles in R' . Foreground area falling outside all the rectangles due to the removal of r_k is denoted as F ;

- Reassign the KLT points to the rectangles in R' and recalculate the new score $S' = \{s'_1, s'_2, \dots, s'_l\}$, $s' = \sum_{l=1}^L s'_l$;

- After the removal of r_k , points in r_k previously is assigned to new rectangles. Denote the distance deviation in r_k as t_k and the distance deviation in new rectangle as t'_k respectively.

(c) *Candidate removal.*

- If $F > 45\%$ of r_k , then keep the candidate;

- elseif $s'' \geq s$ or $\sum_{l=1}^L \delta(s'_l < s_l) \leq 1$, then $R \leftarrow R''$, $s = s''$, $S \leftarrow S''$;

if $s'_l < s_l$, $\delta(s'_l < s_l) = 1$, otherwise, $\delta(s'_l < s_l) = 0$.

- elseif $t'_k \geq t_k + m_1$, then keep the candidate;

- elseif $\sum_{l=1}^L \delta(s'_l < s_l) < N_s^1$ or $t'_k \leq t_k + m_2$,

- then $R \leftarrow R''$, $s = s''$, $S \leftarrow S''$.

(d) $k=k+1$;

K = the number of remaining rectangles in R .

Stage-2: Loop until the rectangles are not changed.

Same as Stage-1 except $\sum_{l=1}^L \delta(s'_l < s_l) < N_s^2 * p_{occ}^k$ in step (c), where

N_s^2 is the minimum number of supporting points for a fully-visible person,

p_{occ}^k is the occlusion percentage of candidate- k in the configuration.

Output:

The number of rectangles, K ; The size and location of each rectangle, $\{x_k, y_k, w_k, h_k\}$, $k = 1, \dots, K$.

Training set: The training images were extracted from the first 300 frames. Twenty images with 79 training persons were used to collect the training patches.

Testing set: The test images are from the remaining 600 frames. Most people are different from those in the training set. One test image is used every five frames, and a total of 119 frames are tested.

The parameters—which were not mentioned in the previous sections—will be listed here. The incremental margins for coherent motion are $m_1 = 15$ and $m_2 = 10$. m_1 should not be too low to cause false alarms, while m_2 should not be too high to cause missed detections.

After getting the foreground mask using an adaptive background estimation method [28], a series of postprocessing steps is performed to create the final foreground region. First, an open operation is performed to remove noises. A structuring element with a radius of one pixel was used in our evaluations. Then, a closing operation is performed to merge broken foreground blobs in the foreground mask. A structuring element with a radius of six pixels was used. Finally, a dilation operation is performed to include almost all the feature points from human beings within the foreground region. In our evaluations, a structuring element with a radius of five pixels was used.

In Fig. 10, the left column shows the foreground region used for each example. It can be observed that it is difficult to get an accurate crowd segmentation based on the foreground contour only. The middle column shows the candidates proposed initially based on head candidates. Since we use a low threshold for candidate proposal, there are usually a number of candidates in the initial configuration. The right column shows the results obtained based on appearance cues. As shown in Fig. 10(b) and (c), although it is not easy to get accurate segmentation based on the foreground contour, some good results have been achieved based on the appearance cues from B-ISM. However, the reliability of appearance cues might be low, particularly when the crowd is dense or the background is complicated. As indicated by the arrows in Fig. 10(a), (c), and (e), some false detections might occur in the final results.

For over 119 test frames, there are 696 persons in the ground truth. The detection rate is 90.66%, and the false-alarm rate is 16.52% in the developed appearance-based method. Results on the USC-Campus sequence have also been reported in [12] and [14]. However, in [14], temporal information from human tracking has been used to handle serious occlusions. Detections in one frame are verified based on the results from the neighboring frames. Since our method is only based on one frame, only the results in [12] will be discussed. The results of the entire sequence in [12] are 92.82% for the detection rate and 0.18% for the false-alarm rate. The good results rely on an informative foreground contour, a detailed 3-D human model, and accurate camera calibration parameters. It is actually assumed that the crowd is not so dense that all the people are visible on the foreground contour. The results of our method are not comparable with [12] yet. However, the segmentation is based on full exploration of appearance cues in the foreground region, which has three advantages. First, it does not require an accurate foreground contour, which might be easily corrupted by noise. A foreground region of the crowd is much easier to obtain, which assures a wide application range of the method. Second, since the segmentation is based on the information inside the foreground area, it is possible to find individuals in a dense area. Third, our method does not require any camera calibration. Camera calibration might not be obtained easily in real situations.

B. Multicue-Based Crowd Segmentation

In the previous section, only the appearance cues have been used. Here, a multicue-based crowd segmentation method is tested using both appearance and motion cues. To demonstrate

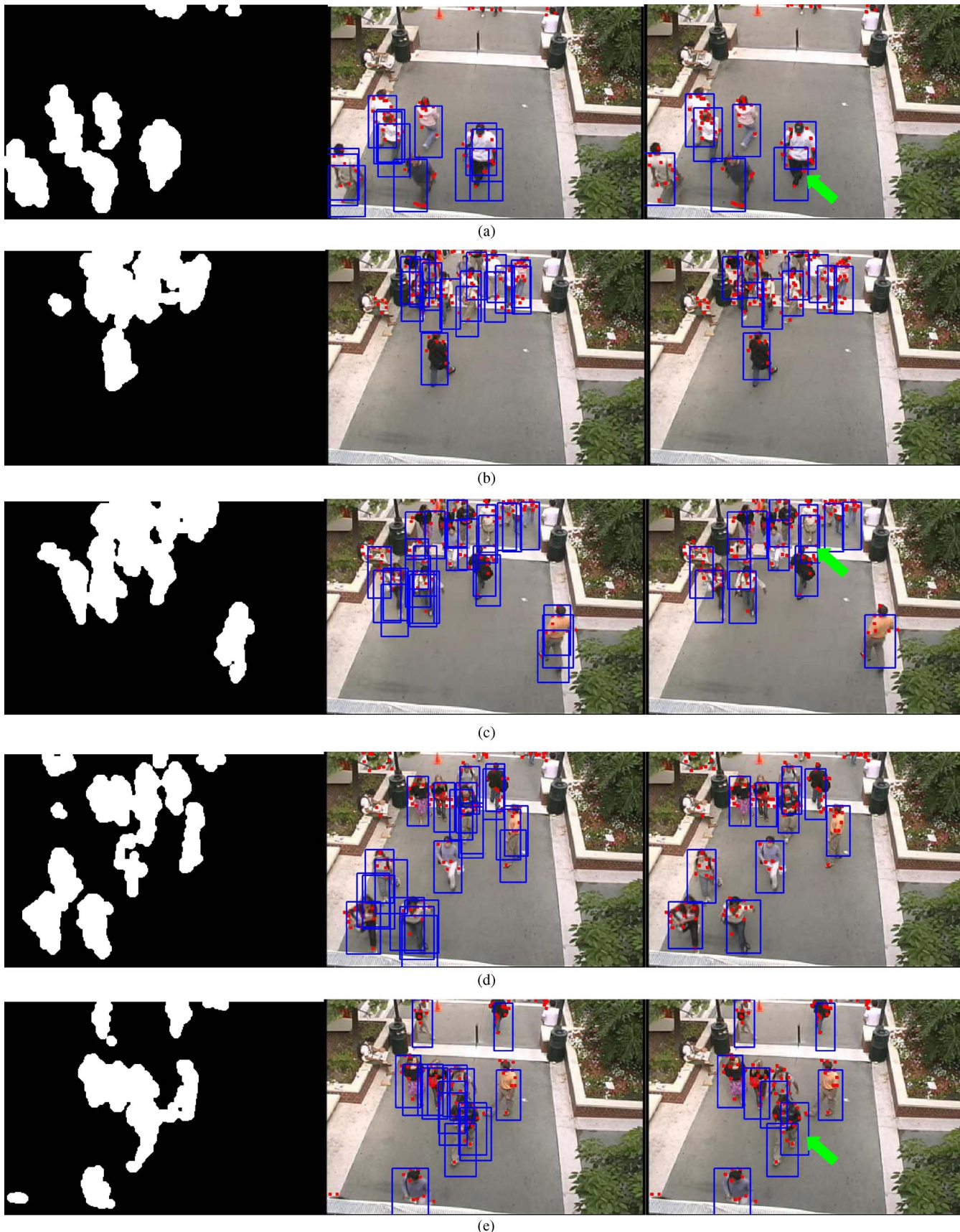


Fig. 10. Selected frames from the USC-Campus sequence. (Left) Foreground region. (Middle) Initial candidates. (Right) Results from the appearance-based method.

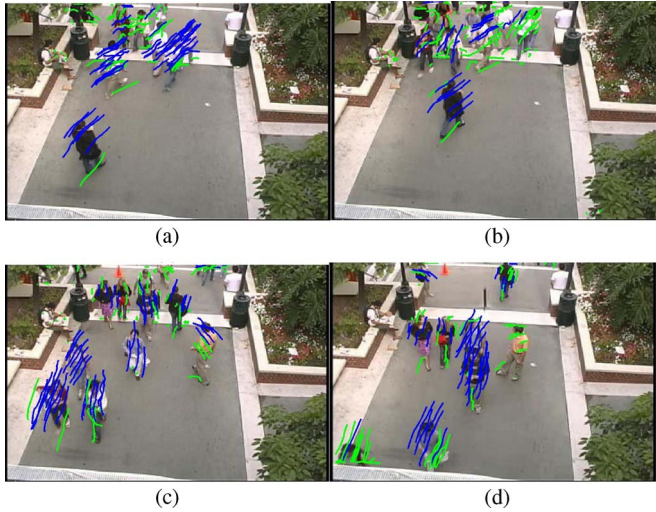


Fig. 11. Sample frames with trajectories of all the points within ± 15 frames. The trajectories of points from a rigid body part are displayed with blue color.

the use of the feature points from rigid body parts, the multicue-based method is tested in two ways: In *Motion-A*, coherent motion is considered based on all the feature points from a candidate, and in *Motion-B*, coherent motion is described based on only the points from the rigid body parts. Results from all the methods are obtained based on more than 119 frames of the USC-Campus sequence.

Before showing the segmentation results of the methods, some sample frames have been shown in Fig. 11 to illustrate the results from detecting points from rigid body parts. The trajectories of all the points along ± 15 frames are displayed. The trajectories of points moving on rigid body parts are highlighted with blue lines. It can be observed that most points from parts with articulated movements (which are mainly from feet) can be excluded. Points from rigid body parts have more consistent trajectories within an individual. A high average standard deviation in distance among all the KLT points in a candidate is more likely due to multiple individuals within it.

In our evaluations, the 2-D ground truth is manually obtained with rectangles. Only fully visible persons are considered for the evaluations. Those whose bodies are partially outside the scene are not counted as matches or errors, i.e., denoted as “do not care.” A detection that has a large overlap ($> 50\%$) with the human object in the ground truth is defined as a correct detection. Each ground truth can have only one correct detection. The detected rectangle without a corresponding person is a false detection.

$$\text{Detection rate} = \frac{(\# \text{correct detection})}{(\# \text{ground truth})}$$

$$\text{False-alarm rate} = \frac{(\# \text{false detection})}{(\# \text{ground truth})}$$

Table III shows the results of the methods. With only the appearance cues, the detection rate and false-alarm rate are 90.66% and 16.52%, respectively. In *Motion-A*, all the points are considered for the measure of coherent motion. It can be observed that the false-alarm rate has been reduced by around

TABLE III
COMPARISON OF THE METHODS

	Detection rate (%)	False alarm rate (%)
Appearance-based	90.66	16.52
<i>Motion-A</i>	89.94	14.80
<i>Motion-B</i>	90.80	14.51

2%. However, the detection rate is reduced by around 1% at the same time. This is because points with articulated movement might also result in incoherent motion, which makes it unreliable to identify multiple persons based on motion incoherence. As a result, some valid candidates might be falsely removed based on the analysis of motion cues. On the other hand, in *Motion-B*, only the points from rigid body parts are considered. As expected, the motion cue has been used more reliably. The false-alarm rate is further reduced than *Motion-A* while the detection rate is increased.

Fig. 12 has shown some sample frames with the results of the appearance-based method, *Motion-A*, and *Motion-B*. Most frames have the same results based on the three methods. The arrows have indicated some cases where improvements have been made by considering motion cues. In Fig. 12(a), (c), and (d), false detections have been removed after the motion cues are considered. In Fig. 12(a), *Motion-A* has got the best result. Due to the points from legs and feet, points in the candidate behind have more consistent trajectories. In Fig. 12(c), the false rectangle covers points from two persons. After the removal of the false candidate, points are assigned to two separate rectangles, in which trajectories are more consistent. In Fig. 12(d), the partially occluded girl has got a more accurate detection with *Motion-B*, in which only points from rigid body parts are used for the measure of coherent motion.

C. More Results From the CAVIAR Data Set

To provide more evaluations of our method, more results based on the CAVIAR video set have been obtained. The CAVIAR data set [29], which includes several video sequences in a corridor, is a commonly used video set for human detection. It was taken by a stationary camera fixed at a few meters above the ground. The image size is 384×288 pixels, and the frame rate is 24 fps.

This set was taken from a different viewing angle from the USC-Campus sequence, and human appearances are quite different. Hence, a new B-ISM is established for this test. In addition, this video set has serious perspective distortions. Perspective correction has been considered for patch extraction in this test.

Training set: The training images were extracted from five video sequences in the CAVIAR data set. Twenty-two images with ten persons were used in our evaluations to form the codebook. Only fully visible persons of a certain size were used. The rough foreground region for the selected training images was obtained manually.

Testing set: The test images are from two video sequences that are different from the training set in the CAVIAR data set.



Fig. 12. Sample frames with results of the appearance-based method, *Motion-A*, and *Motion-B*.

The parameters used in the method are listed here. The threshold for head candidate detection is 0.112 ($\approx 1/9$). All the initial candidates have more than 70% overlap with the foreground area. The minimum number of support points for any candidate in the first stage is three. At Stage-2, the minimum number of support points for a fully visible person is



Fig. 13. More sample frames from video sequences in the CAVIAR set with results of the appearance-based method, *Motion-A*, and *Motion-B*.

five for people close to the camera, and it is three for those far away (y -coordinate < 30). For motion cues, the incremental margins are $m_1 = 25$ and $m_2 = 5$.

Fig. 13 shows some sample results on the CAVIAR video sequences. As highlighted by the arrows in Fig. 13(a), (c), and (e), false detections in the appearance-based method have been removed based on motion cues. However, in Fig. 13(b), a candidate is falsely retained with *Motion-A*. Another candidate is falsely removed with the method *Motion-A* in Fig. 13(d).

V. CONCLUSION AND FUTURE WORK

In this paper, a method based on both appearance and motion cues for crowd segmentation has been presented. The method has formulated crowd segmentation into a feature point clustering process.

With the development of a B-ISM, the information inside a crowd region has been exploited, and our method can detect human individuals in a densely crowded region. In addition, without the requirement of an informative foreground contour, this method can be used in most situations. Coherent motion of feature points from the same individual is combined with appearance cues to achieve better segmentation performance. Three features have been proposed to extract points from rigid human body parts. Results show that coherent motion features can be described more reliably when only points from rigid body parts are considered. Motion cues play an important role when appearance cues become less reliable, such as when the background is cluttered or the crowd is dense. It should be noted that, except for the foreground extraction, the method does not inherently require people to be moving. Significant motion cues are used just as supplementary evidence.

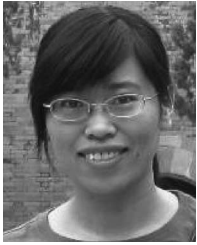
In the future, the following work can be carried out as an improvement of the method. First, instead of a rectangular shape, a more accurate human model can be used. This is good for more accurate localization of feature points and more accurate individual segmentation. Second, more detailed examinations on different interest point detectors and appearance descriptors will be performed, which might improve the performance of the appearance-based method. Third, by performing a further examination of the scores in each body part, an occlusion map could be obtained for the person. The occlusion reasoning results would help remove false detections and handle crowded scenarios.

ACKNOWLEDGMENT

The authors would like to thank T. Zhao for sharing the USC-Campus video sequence.

REFERENCES

- [1] C.-F. Juang and C.-M. Chang, "Human body posture classification by a neural fuzzy network and home care system application," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 37, no. 6, pp. 984–994, Nov. 2007.
- [2] A. Gupta, A. Mittal, and L. S. Davis, "Constraint integration for efficient multiview pose estimation with self-occlusions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 493–506, Mar. 2008.
- [3] C.-F. Juang, C.-M. Chang, J.-R. Wu, and D. Lee, "Computer vision-based human body segmentation and posture estimation," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 39, no. 1, pp. 119–133, Jan. 2009.
- [4] J.-W. Hsieh, C.-H. Chuang, S.-Y. Chen, C.-C. Chen, and K.-C. Fan, "Segmentation of human body parts using deformable triangulation," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 40, no. 3, pp. 596–610, May 2010.
- [5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.
- [6] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1491–1498.
- [7] O. Tuzel, F. Porikli, and P. Meer, "Human detection via classification on Riemannian manifolds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
- [8] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 32–39.
- [9] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 734–741.
- [10] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 7–13.
- [11] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.
- [12] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2003, pp. 459–466.
- [13] T. Zhao and R. Nevatia, "Tracking multiple humans in complex situations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1208–1221, Sep. 2004.
- [14] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1198–1211, Jul. 2008.
- [15] J. Rittscher, P. H. Tu, and N. Krahnstoeber, "Simultaneous estimation of segmentation and shape," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 486–493.
- [16] Y.-L. Hou and G. K. H. Pang, "People counting and human detection in a challenging situation," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 41, no. 1, pp. 24–33, Jan. 2010.
- [17] L. Dong, V. Parameswaran, V. Ramesh, and I. Zoghliami, "Fast crowd segmentation using shape indexing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
- [18] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *Proc. Workshop Stat. Learn. Comput. Vis., ECCV*, 2004, pp. 17–32.
- [19] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 878–885.
- [20] G. J. Brostow and R. Cipolla, "Unsupervised Bayesian detection of independent motion in crowds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 594–601.
- [21] V. Rabaud and S. Belongie, "Counting crowded moving objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 705–711.
- [22] E. Seemann, B. Leibe, K. Mikolajczyk, and B. Schiele, "An evaluation of local shape-based features for pedestrian detection," in *Proc. Brit. Mach. Vis. Conf.*, 2005, pp. 11–20.
- [23] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Int. Joint Conf. Artif. Intell.*, 1981, pp. 674–679.
- [24] J. Shi and C. Tomasi, "Good features to track," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1994, pp. 593–600.
- [25] Y.-L. Hou and G. K. H. Pang, "Human detection in a challenging situation," in *Proc. IEEE Int. Conf. Image Process.*, 2009, pp. 2561–2564.
- [26] O. Arandjelovic and A. Zisserman, "Crowd detection from still images," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 525–532.
- [27] S. Birchfield, Source Code of the KLT Feature Tracker, 2006. [Online]. Available: <http://www.ces.clemson.edu/~stb/klf/>
- [28] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, Aug. 2000.
- [29] [Online]. Available: <http://homepages.inf.ed.ac.uk/rbf/CAVIAR/>



Ya-Li Hou received the B.Eng. degree in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, in 2004, the M.S. degree from Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai, China, in 2007, and the Ph.D. degree from The University of Hong Kong, Pokfulam, Hong Kong, in 2011.

She is currently with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China. Her research interests include visual surveillance, pattern recognition, and computer vision.



Grantham K. H. Pang (S'84–M'86–SM'01) received the Ph.D. degree from the University of Cambridge, Cambridge, U.K., in 1986.

From 1986 to 1996, he was with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. He is currently an Associate Professor with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam, Hong Kong. He acts as a Consultant to a number of local and international companies and has served as an expert witness for the Courts of the Hong Kong Special Administrative Region. He has published more than 160 technical papers and has authored or coauthored six books. He is the holder of five U.S. patents. His research interests include machine vision for surface defect detection, optical communications, expert systems for control system design, intelligent control, and intelligent transportation systems.