Available online at www.sciencedirect.com

ScienceDirect

Pattern Recognition 41 (2008) 2447–2460

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/pr

Tracking multiple objects through occlusion with online sampling and position estimation

Lin Zhu, Jie Zhou*, Jingyan Song

Department of Automation, Tsinghua University, Beijing 100084, PR China

Received 25 October 2006; received in revised form 29 November 2007; accepted 16 January 2008

Abstract

To track multiple objects through occlusion, either depth information of the scene or prior models of the objects such as spatial models and smooth/predictable motion models are usually assumed before tracking. When these assumptions are unreasonable, the tracker may fail. To overcome this limitation, we propose a novel online sample based framework, inspired by the fact that the corresponding local parts of objects in sequential frames are always similar in the local color and texture features and spatial features relative to the centers of objects. Experimental results illustrate that the proposed approach works robustly under difficult and complex conditions.

Keywords: Multiple objects tracking; Occlusion; Online sampling; Position estimation

1. Introduction

In video surveillance systems, accurate and real-time multiple objects tracking will greatly improve the performance of objects recognition, activity analysis and high level event understanding [1–5]. Segmentation and tracking multiple objects is important not only for visual surveillance, but also for other video analysis applications such as video indexing, video archival and retrieval systems [6]. Many video surveillance systems are required to keep tracking multiple objects (persons and vehicles) as they pass through the scene even when they are occluded by or interacting with other moving objects in the scene. In these cases, it is quite possible for the trackers to miss the objects. For real applications, it is even more important to track multiple objects when they are interacting than when they are isolated [6]. Multiple objects tracking, especially tracking and segmentation of multiple objects with occlusion, is a challenging problem for the current research.

Many researches aim to handle the problem of tracking multiple objects with occlusion. Such approaches for tracking with/without occlusion often assume that the objects are moving smoothly, and then the filter based models are widely used for predicting the position of objects during occlusion. But in actual scenes, objects may move irregularly. Then the false prediction of position may cause tracking error and the accumulation of error may even make the tracking failure [7,8]. Feature correspondences between sequential frames are also widely used in tracking systems. Some tracking approaches are based on the objects' global features, such as size, color, shape, velocity and center of mass. These approaches can hardly solve the problem of tracking multiple objects through occlusion. Some other approaches rely on the detection of local parts of moving objects, for example, the corners of the cars. These approaches can partially solve the problem of tracking through occlusion. But it is a difficult problem to decide which object each part belongs to [9]. Some tracking approaches are based on the parametric spatial models of the objects [10–14], and can track objects with partial occlusion. But those approaches rely on the prior parametric spatial models of objects and the computational cost of parameter estimation is usually high for iteration. If the objects are not the supposed kind or the poses of the objects do not meet the supposed pose (upright pose), the tracking will fail. In these approaches, in order to track multiple objects through occlusion, a lot of prior information is

* Corresponding author. Tel./fax: +86 10 62796881.
E-mail addresses: zhul01@mails.tsinghua.edu.cn lin.zhu.612@gmail.com (L. Zhu), jzhou@tsinghua.edu.cn (J. Zhou), jysong@tsinghua.edu.cn (J. Song).

0031-3203/$30.00 © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.01.014

assumed known, such as the ground plane constraint [15–17], the kind of objects (to build the 2D or 3D special model before tracking) [10–14], certain pose (upright pose) [14] or objects' smooth motion models [10,11,15,18–21].

Recently, appearance model based tracking approaches have drawn more and more attention.

Appearance model based tracking is popular for region tracking. Jepson et al. [22] presented a Gaussian mixture model for motion based tracking, combining the stable region structure with two-frame information and an outlier model. Nguyen et al. [23] proposed a method for tracking a region based on appearance rather than geometry, and assumed that all the object points have similar motion. These region tracking methods based on appearance model rely on the smooth motion assumption, and will fail in most situations in which the object becomes entirely occluded for a long time.

Appearance models are also very useful to object modeling for multiple objects tracking. Wren et al. [24] represented a person as a set of blobs representing the major parts of the body as the torso, bottom and head, which was used to track single person in the scene. Elgammal and Davis [6] used a similar model to deal with the problem of tracking multiple persons through occlusion. A person in an upright pose was modeled as a set of vertically aligned blobs, and then each pixel of foreground region was classified using the maximum likelihood classification. This approach does not rely on the ground plane constraint and can learn the persons' appearance information before occlusion to classify the persons' region through occlusion. But the approach in Ref. [6] relies on the assumptions of upright pose and slight position changes between consecutive frames, which may not be reasonable in real cases. When we are not sure of the kinds of objects before tracking, the approaches above are unsuitable. Furthermore, the method in Ref. [6] has a heavy computational cost. To reduce this cost, Hu et al. [25] made some simplification, but the method still relies on the prior person spatial model and upright pose assumption.

In this paper, a novel online sampling based approach is presented to solve the problem of segmentation and tracking multiple objects through occlusion for static cameras without the constraints of object's kind, pose, smooth motion and ground plane. The new approach performs online extraction of the training samples from the objects' isolated regions before occlusion. And the new approach trains online to classify the samples extracted from the foreground region in the following frames through occlusion.

Our approach is inspired by the observation that, for local parts of one object in current frame, there always exist corresponding local parts with similar texture feature and spatial feature relative to the center of the object in consecutive frames, as illustrated by the blue and green local parts in Fig. 1. Due to deformation, some corresponding local parts with similar texture feature may have some difference in spatial feature or may become invisible in some frames, like the blue part. For this type of local parts, it is possible to find some local parts in some previous frames with the similar texture and relative spatial features. For other type of local parts in current frame, the corresponding local parts with similar texture feature in previous frames are very stable in spatial feature relative to the center of object, like the green part in Fig. 1. These parts are considered as spatial distinguishing parts, which can be used to estimate the position of the center of object in current frame.

Contrast to the approaches mentioned above (e.g. Ref. [6]), the new approach has the following characteristics:

- Block sample based instantaneous object representation: The objects' local parts are extracted by window shifting to get the samples. Each sample is described by the local color and texture features and relative spatial features to the objects' centers. Then the instantaneous spatial features of the objects are presented without the prior fixed spatial models of the objects. In Ref. [6], all the pixels in one blob are assumed to have the same color density and the blob spatial models are fixed before occlusion. The representation in Ref. [6] did not consider the differences between pixels caused by local texture and spatial location relative to the objects' centers.
- Online learning and classification scheme: The segmentation problem of the foreground region through occlusion is transferred to the online classification problem of the testing samples extracted from the foreground region. Training samples are extracted from the instantaneous objects' regions in previous frames before occlusion, and the foreground regions through occlusion in following frames must be classified into multiple objects. So, the number of training samples is quite small and the training and classification process must be accomplished very quickly. Since the samples extracted from different parts of an object have the same labels, the special online classification scheme based on nearest neighborhood classification is proposed to solve this problem effectively and quickly.

Based on the above characteristics, the proposed approach has the following advantages:

- To distinguish similar objects better, the appearance based object model is improved by adding the texture features and instantaneous spatial features relative to the center of object.
- By extracting instantaneous spatial features rather than fixed prior spatial models, different kinds of objects can be simultaneously tracked efficiently. Previous model based approaches usually supposed the kind of tracking objects and built a parametric prior spatial model. These models are not fit for other kinds of objects, and are easy to make mistakes when the person is not in the supposed pose.
- It can deal with the cases of frame skipping or unsmooth position changing between consecutive frames. In reality, the position of the same object in consecutive frame may change unsmoothly, because of frame skipping by limited net transmission or large pose changing. To get the exact position of origin, many traditional approaches assume that there are only slight or smooth position changes between consecutive frames. Some approaches use the position of previous frame to replace the position of current frame with the assumption of position changing slightly. Some other approaches predict the position with special filter models with the assumption of

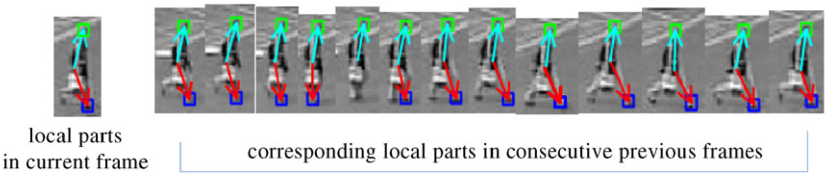
local parts in current frame
corresponding local parts in consecutive previous frames

Fig. 1. For local part of one object, two types of similarities of the texture feature and spatial feature relative to the center of the object in consecutive frame.

position changing smoothly. When the slight/smooth movement assumption is unreasonable, the approach may fail. Our approach can estimate the center of objects robustly through occlusion without these assumptions. So, our approach can be combined with the traditional methods to reduce the tracking error caused by position mis-estimation.

The paper is organized as follows. In Section 2 we explain our approach in detail. In Section 2.1, the proposed framework is introduced. From Sections 2.2 to 2.8, we describe respectively the steps of occlusion judgment, sample extraction, online updating training samples, online learning, objects' centers estimation, label classification and post-processing. In Section 3, experiments are carried out and the results demonstrate the favorable performance. The paper concludes in Section 4.

2. Method description

2.1. System framework

In some complex cases, the ground plane is invisible, or the prior spatial model or motion model is not known before tracking. So, we cannot rely on the related assumptions to track multiple objects through occlusion. We represent a novel online sample based approach to deal with this problem using the local color and texture features and relative spatial features from several frames before occlusion. The system flowchart is shown in Fig. 2. The processing steps are explained briefly in the following.

When one frame comes, the foreground regions corresponding to multiple objects can be extracted using a background subtraction algorithm [24]. With the dilation/erosion operators, the fragments of one object can be connected and the noises are eliminated. Some parts of objects which are similar to the background may be missed, but these can hardly affect the results of tracking through occlusion. Before occlusion occurs, many methods can track multiple objects well. The tracking results of the objects before occlusion are used for online updating the training samples of these objects. The tracking system needs the occlusion judgment module to judge if occlusion occurs. When the foreground regions corresponding to multiple isolated objects are merged into one foreground, occlusion is considered to occur. The problem we need to solve is to correctly segment the foreground region into multiple objects through occlusion.

In this paper, we try to use the online sample based classification approach to solve the problem of tracking through occlusion. The region information of multiple objects is known before occlusion. By block sampling, the instantaneous local color and texture features and the instantaneous relative spatial features of these regions are extracted to form the training samples with labels. The foreground regions in following frames must be classified into multiple objects. Here we face two difficulties: The first one is that the number of training samples for learning is rather small. The second one is that the classification approach must be fast and effective to meet the online classification requirement. Traditional sample based classification framework can hardly deal with this problem.

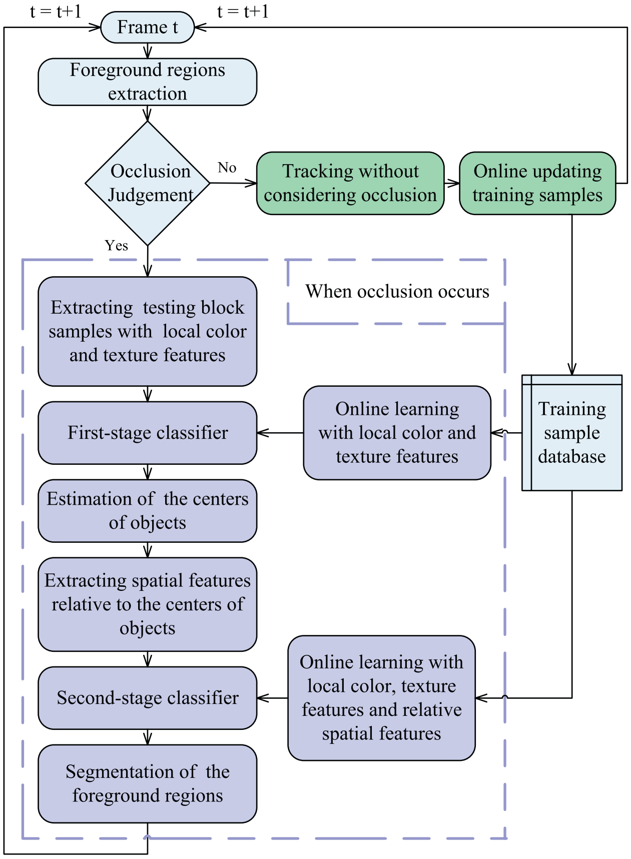
t = t+1
Frame t
t = t+1
Foreground regions extraction
Occlusion Judgement
No
Tracking without considering occlusion
Online updating training samples
Yes
When occlusion occurs
Extracting testing block samples with local color and texture features
First-stage classifier
Online learning with local color and texture features
Training sample database
Estimation of the centers of objects
Extracting spatial features relative to the centers of objects
Second-stage classifier
Online learning with local color, texture features and relative spatial features
Segmentation of the foreground regions

Fig. 2. The system's flowchart of the tracking process.

Considering that the local features of objects are still visible during occlusion, also that the local textures of the same object between consecutive frames are similar (even when the objects have greatly global deformation), the new approach extracts the labeled training samples by block sampling from the foreground regions of objects in several frames before occlusion. The samples are significantly different from the traditional sample based classification approach, as all the samples of one object have the same label, though they may represent different parts of the object (for example, one sample represents head, the other represents feet). Also, the samples are special in the sense that samples of different objects have the different labels, though they may represent the same parts of different objects (for example, they represent the face part of different objects).

When occlusion occurs, the training samples are online extracted from the labeled foreground regions of several previous frames before occlusion. Obviously, the number of training samples will be very small. Considering the sparsity and the speciality of the training samples, an appropriate online classification method is needed. For our system, K nearest neighborhood classification is used to classify the sample blocks from the foreground region corresponding to multiple objects of current frames. In order to use relative spatial features without smooth motion assumption, we use first-stage classifier to estimate the position of the centers of objects more accurately. With the estimated centers of objects of current frame, the second-stage classifier is obtained with the relative spatial features and the local color and texture features.

Then, with post-processing to the classification results, the foreground regions are segmented into multiple objects. When the foreground region corresponding to multiple objects is split into multiple regions, occlusion is considered to end.

By block-sampling the objects' regions and novel position estimation, the new approach can track multiple objects through occlusion, even with severe deformation, large pose changes and unsmooth motion. Obviously, the new approach can combine with other approaches (if reasonable) to get more accurate segmentation and tracking results. The detailed steps are explained in the following.

2.2. Occlusion judgment

Many tracking algorithms can judge the beginning and the end of occlusion, for example, the bounding box distance measure [12] or the connected component method [26]. In this paper, with the results of background subtraction, multiple connected foreground component can be obtained after removing small components. The correspondence between the tracks in previous frames and the foreground regions in current frames are obtained by judging the shortest Euclidean distance between them. To handle the merging and splitting of multiple regions, several rules are used as follows.

- If two or more tracks in previous frame correspond to one foreground region in current frame, multiple regions are considered to be merged, and the occlusion is considered to occur.
- If occlusion was not considered to occur in previous frame, and a single track in previous frame corresponds to two or more foreground regions in current frame, all these regions are processed as one object, and these foreground regions are considered as fragments. The regions should be separated into two or more objects and a new track should be created only if none of foreground regions corresponds to the track.
- If occlusion was considered to occur in previous frame, and multiple tracks in previous frames correspond to more than one foreground regions in current frame, all these foreground regions are processed with proposed method for segmentation through occlusion. The occlusion is regarded to end until the label results of pixels are uniform within each foreground region and different from other regions.

2.3. Extracting training/testing samples

For sample based classification, the first thing to be considered is how to get the training samples. We assume that before occlusion, the isolated objects are tracked well, and the foreground regions are all labeled correctly. The objects' regions in n frames before occlusion are kept and divided into many blocks as training samples. The sample blocks are obtained using certain size square window shifting on the foreground region horizontally and vertically. If the area of foreground region in the window is smaller than 50% of the area of the window, the sample will be thrown away. To obtain no less than 100 samples for each object, the width of window L and the number of frames n are confirmed according to the objects region in the scene. The width of the window is between 15 and 25 pixels in experiments, and the frame number n is between 5 and 15 frames (3 frames at least). The window width is L pixels, and the shift step is set to be $L/2$ pixels. If no larger than 10 samples can be extracted from the smallest object region in one frame, the window width L will be decreased and the frame number n will be increased. Sometimes there is great disparity in area between the objects under tracking. If the number of pixels of the smallest region among the multiple objects is set to S, the window width L will be adaptive as according to the equation as follows:

$$L = \begin{cases} \sqrt{S/5} & \text{if } \sqrt{S/5} < 15, \\ 15 & \text{if } 15 \leqslant \sqrt{S/5} < 25, \\ 25 & \text{if } 25 \leqslant \sqrt{S/5}. \end{cases}$$

For example, if a person is much smaller than a car as shown in Fig. 3(a) (from the dataset1 (CAM 1) for PETS'2001 [27]), where the window width is 15 pixels. The region of the smallest object (people) can only be divided into less than 10 blocks, then the number of the frame n should be larger ($n > 10$) for person. While for the big object (vehicle), there are almost 100 blocks in one frame, as shown in Fig. 3(b). It is enough for n to be 3 for vehicle.

As a result, for each object there are over 100 sample blocks, which are used as the training samples of supervised learning. If the width of the sample window is L, the number of pixels for each sample is $p = L \times L$. We transform the p pixels as a vector $V_{train} = [V_1, \ldots, V_p]$ to describe the local color and texture features of the sample, and the spatial distance from the center point of the block to the center of the object region is $\vec{D} = [\vec{D}_x, \vec{D}_y]$ to describe the relative spatial features of the sample.

For a current frame through occlusion, the testing samples can also be extracted with the same window width and sift step. The local color and texture features of the testing samples can

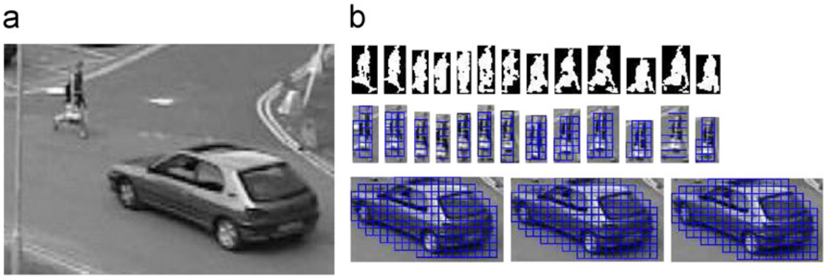

Fig. 3. (a) One frame of a person who is going to pass by a car. (b) The sampling results. Top: Templates of the isolated person regions for frame 540 to frame 552. Middle: Samples of person for frame 540 to frame 552. Bottom: Samples of car for frame 550 to frame 552.

also be transformed as a vector $V_{test} = [V_1, \ldots, V_p]$. Since the positions of the centers of multiple objects are unknown, we can only obtain the absolute position of the center of the ith block denoted as $\vec{P}_i$.

We consider all the sample blocks of two objects as training samples, without considering the difference of the blocks labeled as the same objects. This is different from the traditional training sample. For the two blocks belonging to the same object, even if one is from the head and the other is from the leg, the labels of the two blocks are the same. With these labeled training samples many supervised learning methods can be used to produce a set of classifiers.

2.4. Online updating training samples

Before occlusion, the objects' tracking results of nearest previous m frames are stored (m is 20 in our system). These results can be obtained from other tracking approaches without occlusion. If occlusion does not occur in current frame, the foreground regions of the objects are saved, and that of oldest previous frame are thrown away. The foreground regions are online updated frame by frame. When occlusion occurs, the training samples are extracted from the n nearest frames before occlusion. Then, classifiers are learned from these online training samples. Note that only hundreds of training samples from previous frames are used to train classifiers. So, the computational cost of this online sample based classifier training approach is not high.

2.5. Learning method

Though the global shape and appearance of one object may change greatly in consecutive frames due to deformation or occlusion, there is still large similarity in local color and texture features and relative spatial features between testing samples and training samples after dividing the global of object to small blocks. Due to the limited online training samples, the time requirement of online classification and the particularity of the samples, the K nearest neighborhood method is used in our system. Firstly, local color and texture features are used for first-stage classifier to select some distinguishing center points of the samples. Using spatial features of these points and their corresponding similar testing samples, the positions of the centers of the objects are estimated. Secondly, with the estimated centers' positions, the local color and texture features and the relative spatial features are used to classify the testing samples more accurately. From the experimental results, we can see the K nearest neighborhood method can get good result in complex cases.

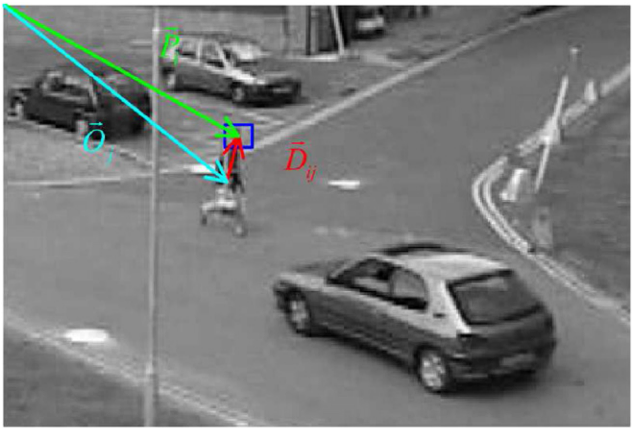

Fig. 4. One training sample's instantaneous spatial feature.

2.6. Estimating objects' centers

In order to get the relative spatial features of these testing samples, the centers of multiple objects in current frame need to be estimated correctly. Considering the possibility of large changes of positions between consecutive frames, traditional position estimation methods based on motion smoothness may fail. A new method based on the similarity of local color and texture features is presented to estimate the centers of multiple objects through occlusion without the assumption of smooth motion.

Before estimation, we explain some variables. The distance from ith point to the center of the class j region is denoted as $\vec{D}_{ij}$. We name the coordinate of one point relative to the top-left corner of the frame as the absolute coordinate of the point. Then, the absolute coordinate of the ith point is denoted as $\vec{P}_i$. The absolute coordinate of the center of the class j region which the ith point belongs to is denoted as $\vec{O}_j$. Obviously, $\vec{P}_i = \vec{O}_j + \vec{D}_{ij}$, as shown in Fig. 4.

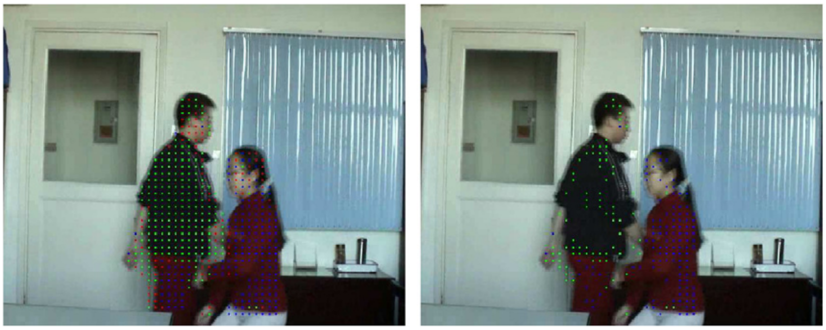

Fig. 5. Left: The label result in frame 34 after first-stage classification. Right: The selected points in frame 34.

When the current frame comes (here we consider only two objects), if the occlusion is considered to occur by occlusion judgment module, the foreground region corresponding to two objects is divided into sample blocks using the forenamed sample extracting method. For all the center points of training samples in previous frames, $\vec{P_i}$ and $\vec{O}_j$ are known. So, $\vec{D}_{ij}=\vec{P_i}-\vec{O}_j$. For current frame, in order to estimate $\vec{O}_j$, $\vec{P_i}$ and $\vec{D}_{ij}$ must be obtained for the center points of testing samples. The details of this procedure are given as follows.

Firstly, we get some points from the foreground region corresponding to multiple objects in current frame. There are many methods to define the points. Here, the center points of each sample blocks using the forenamed sample method are used to select out some spatial distinguishing points. The features of each point in current frame obtained are the local color and texture features of the sample block as vector $V_{test}=[T_1,\ldots,T_p]$, and the absolute positions of the ith point as vector $\vec{P_i}$.

Secondly, for each point, similar center points from all the center points of training samples are found only using the local color and texture features with K nearest neighborhood classification method. This is named first-stage classification. The similarity measurement function for first-stage K nearest neighborhood classifier is $f_1=\|(V_{train}-V_{test})\|$.

To explain more clearly, we simplify the number of kinds to two (with the label L_1 and L_2) and use $K=3$ nearest neighborhood classifier. For the ith point of current frame, according to the distance defined by f_1, the nearest three center points of training samples are i_1, i_2, i_3. Among these three points, the number of points belonging to label L_1 is N_{i1}, the number of points belonging to label L_2 is N_{i2}. All the points are classified to get the membership that describes how much the ith point belongs to each class, denoted as A_i. $A_i=\{(a_{i1},a_{i2})|a_{i1}+a_{i2}=1, a_{i1}>=0, a_{i2}>=0\}$, where a_{ij} is the probability of the ith point belonging to the class j. If using the $K=3$ nearest neighborhood classifier, $a_{ij}\in\{0, 0.333, 0.666, 1\}\cdot a_{i1}=N_{i1}/3$, $a_{i2}=N_{i2}/3$, $N_{i1}+N_{i2}=3$. If $A_i=(1,0)$, the ith point is labeled class 1 and it is filled by blue; if $A_i=(0,1)$, the point is labeled class 2 and it is filled by green. If $A_i=(0.333, 0.666)$ or $A_i=(0.666, 0.333)$, the label of the point cannot be judged. We fill the point by red. The label result is shown in the left of Fig. 5.

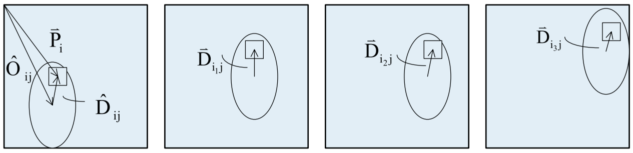

Fig. 6. Relative spatial information for the ith point.

Thirdly, the spatial distinguishing points are selected from the results of first-stage classifier. We suppose that the ith point in current frame is labeled class j in the above step. The corresponding three nearest neighborhood center points of training samples belonging to class j are the i_1th, i_2th, i_3th points, with the relative distance from these points to the region center of the class j $\vec{D}_{i_1j}$, $\vec{D}_{i_2j}$ and $\vec{D}_{i_3j}$ in previous frames. If the three distance vectors $\vec{D}_{i_1j}$, $\vec{D}_{i_2j}$ and $\vec{D}_{i_3j}$ are quite consistent, the ith point in current frame is considered more stable in space. So, the ith point in current frame is selected as spatial distinguishing point; else, it is discarded away. The remaining points are shown in the right of Fig. 5.

Fourthly, the centers of each objects in current frame are estimated with the remained points in above steps.

For the ith point in current frame, we can estimate the relative distance of the ith point to the unknown region center of the class j, $\hat{D}_{ij}=(\vec{D}_{i_1j}+\vec{D}_{i_2j}+\vec{D}_{i_3j})/3$. Then, for the ith testing sample labeled class j, the coordinates of region center of class j in current frame can be estimated by $\hat{O}_{ij}=\vec{P_i}-\hat{D}_{ij}$, as shown in Fig. 6.

For all the points judged to class j in the current frame, we can average the $\hat{O}_{ij}$ to get the absolute coordinate of region center of class j in current frame $\hat{O}_j$.

$$\hat{O}_j=\frac{\sum_{i=1}^{N}a_{ij}\hat{O}_{ij}}{\sum_{i=1}^{N}a_{ij}} \quad \text{for } a_{ij}\in\{0,1\}.$$

Also, we can get the coordinate of region centers for other classes. The results of estimated centers are shown in Fig. 7. Though the estimated centers may still have some difference from the real centers' position, the experimental results are shown that if the error of the centers' estimated position is in some range, the influence to the final labeling error is small.

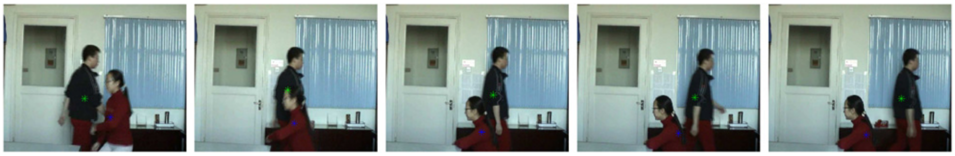

Fig. 7. The estimated centers of objects through occlusion.

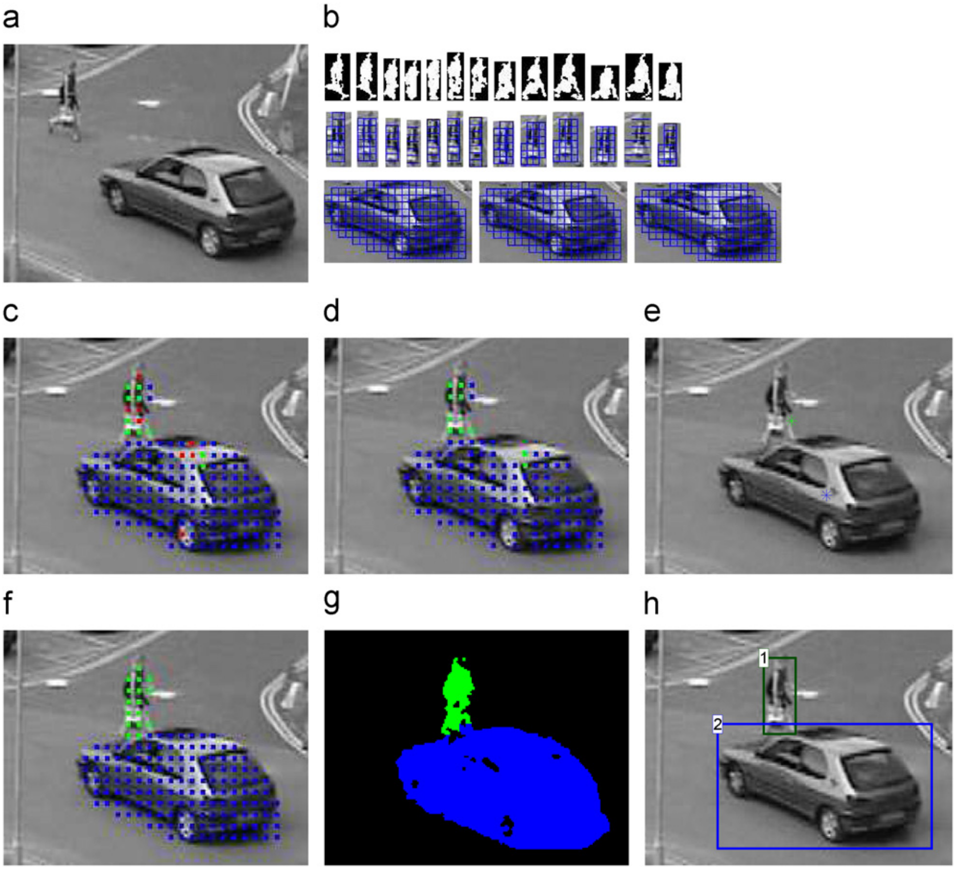

Fig. 8. (a) One frame of a person who is going to pass by a car (frame 540). (b) The sampling results. Top: Templates of the isolated person regions for frame 540 to frame 552. Middle: Samples of person for frame 540 to frame 552. Bottom: Samples of car for frame 550 to frame 552. (c) The first-stage classification result of frame 559. (d) The selected first-stage classification result of frame 559. (e) The estimated coordinates of center of every object. (f) The second-stage classification results of frame 559. (g) The pixel label results of frame 559. (h) The tracking result of frame 559 with labeled bounding box.

2.7. *Label classification*

With the estimated positions $\hat{O}_j$ of region center for every class j, we can calculate the relative distance of ith testing sample's center to the estimated position of class j region center for ith point, $\hat{D}_{ij} = \vec{P}_i - \hat{O}_j$.

For the second-stage classification, with the estimated positions of the centers, we train the classifiers again with the local color and texture information and the relative spatial information. The similarity measurement function for K nearest neighborhood classifier is $f_2 = \|(V_{train} - V_{test})\| + W * \|(D_{train} - D_{test})\|$. All the testing sample blocks are classified to get the membership that describes how much the sample belongs to each class, denoted as B. If the number of kinds is 2, for the ith testing sample, then $B_i = \{(b_{i1}, b_{i2}) | b_{i1} + b_{i2} = 1, b_{i1} >= 0, b_{i2} >= 0\}$, where b_{ij} is the probability of the ith testing sample belonging to the class j. If using the $K = 3$ nearest neighborhood classifier, $b_{ij} \in \{0, 0.333, 0.666, 1\}$. If $B_i = (1, 0)$, the sample is judged as belonging to class 1 and its center point is filled by blue; if $B_i = (0, 1)$, the sample is judged as belonging to class 2 and its center point is filled by green. If $B_i = (0.333, 0.666)$ or $B_i = (0.666, 0.333)$, the label of the sample cannot be judged. We fill the center of this sample by red.

2.8. *Post-processing*

After classifying, labels of some blocks are not determined. According to the label results of the neighborhood blocks, we reconsider the labels of these blocks. If the number of blue

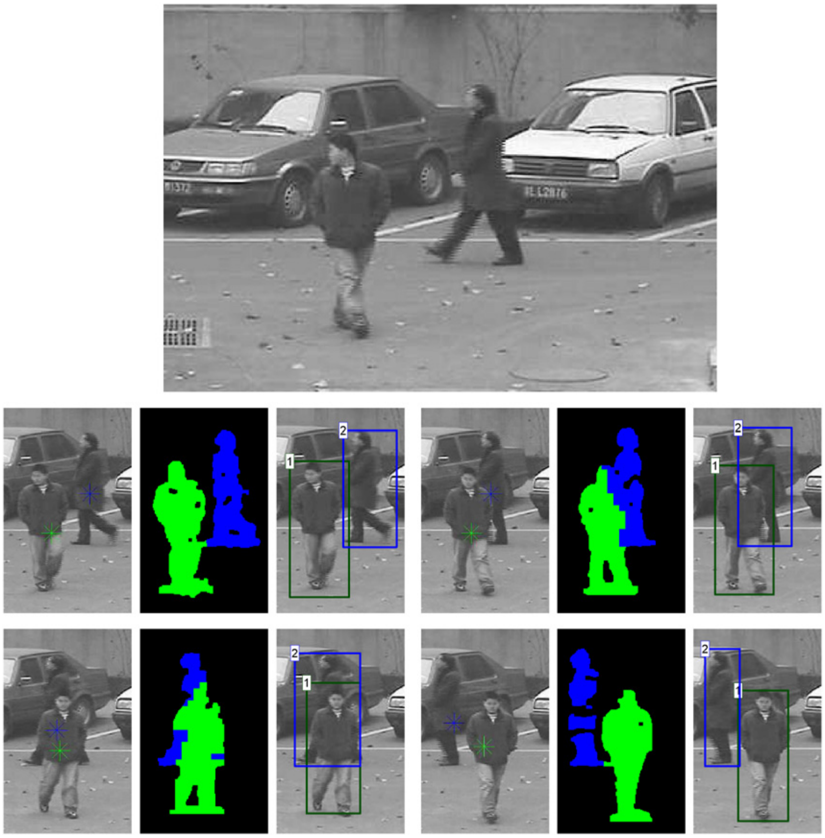

Fig. 9. The first is one frame before occlusion. The others are sample label results for key frames. Left: The estimated centers of the objects. Middle: The foreground region segmenting result. Right: The estimated centers of objects and the final tracking result. The frame numbers from the second row to the third row are respectively 364, 370, 376, 385.

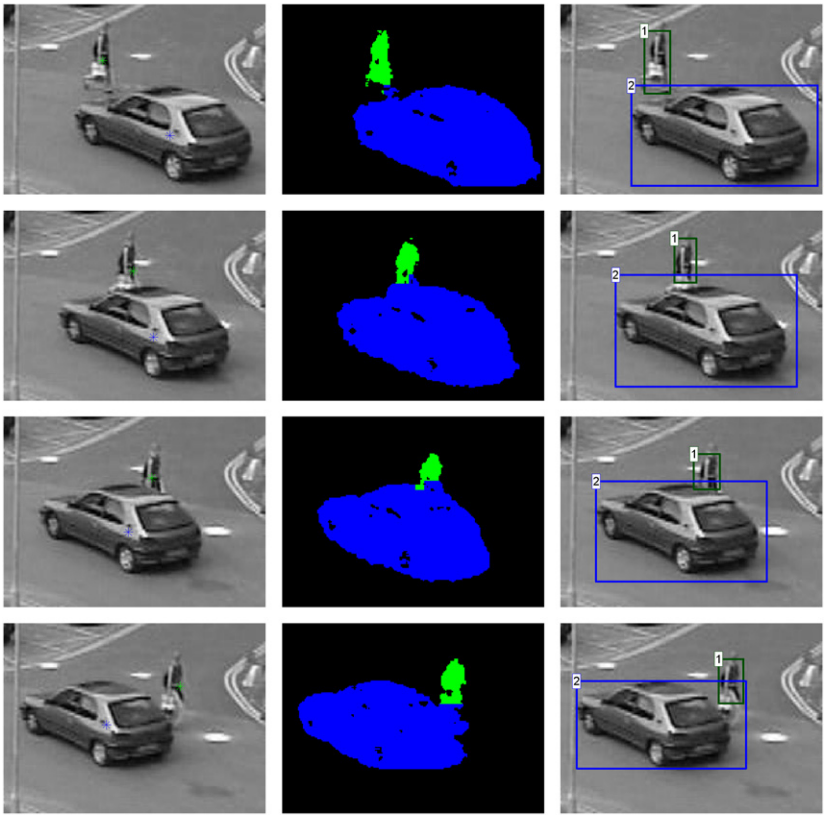

Fig. 10. Experimental results: The frame numbers from the left to the right are respectively 553, 563, 573, 582. The color points are estimated positions of centers of objects.

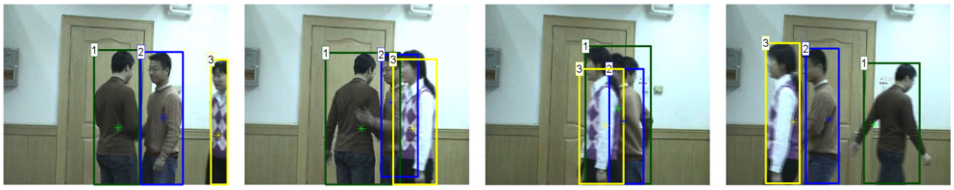

Fig. 11. Experimental results: The frame numbers from the left to the right are respectively 50, 80, 98, 112.

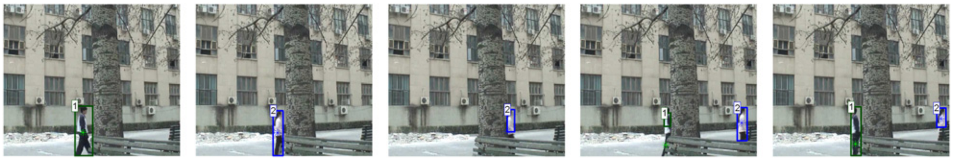

Fig. 12. Experimental results: The frame numbers from the left to the right are respectively 121, 252, 277, 304, 309.

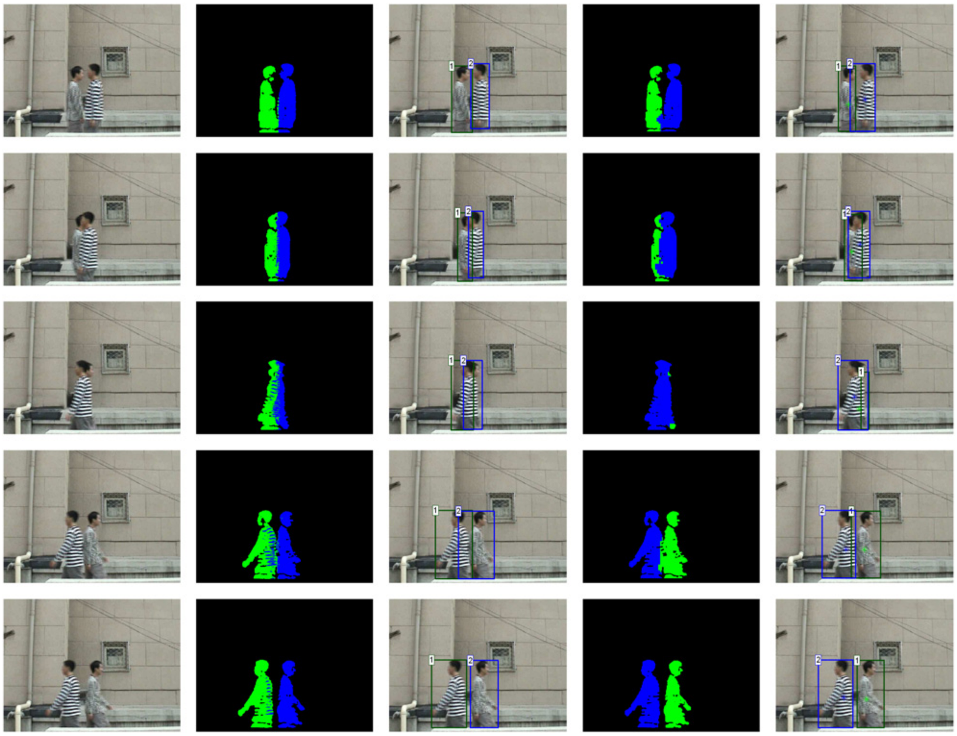

Fig. 13. Comparison results. Left column: The original image. The two columns in the middle: The results using the method of Elgammal and Davis [6]. The two columns in the right: The results using our new approach. (The frame numbers from the top to the bottom are respectively 27, 28, 29, 30, 31.)

points or green points in its 8 neighborhood is larger than the number of other color points of all the valid neighborhood points, the block is relabeled to that label. On the other hand, for green points and blue points, if the point is different from the points around it, its color will be relabeled to make the labels more smooth. At last, with the block label results, the foreground region is filled in pixels. With this result, we can make many judgments, for example occlusion depth order, occlusion degree and the correspondence relations of foreground.

In Fig. 8, the result of every step of our framework is shown. (a) Shows one frame of a person who is going to pass by a car (frame 540). (b) Shows the sampling results. Top: Templates of the isolated person regions for frame 540 to frame 552. Middle: Samples of person for frame 540 to frame 552. Bottom: Samples of car for frame 550 to frame 552. (c) Shows the first-stage classification result of frame 559. (d) Shows the selected first-stage classification result of frame 559. (e) Shows the estimated coordinates of center of every object. (f) Shows the second-stage classification results of frame 559. (g) Shows the pixel label results of frame 559. (h) Shows the tracking result of frame 559 with labeled bounding box.

3. Experimental results

To validate the effectiveness of our new approach in complex cases, we select experimental data from some typical cases.

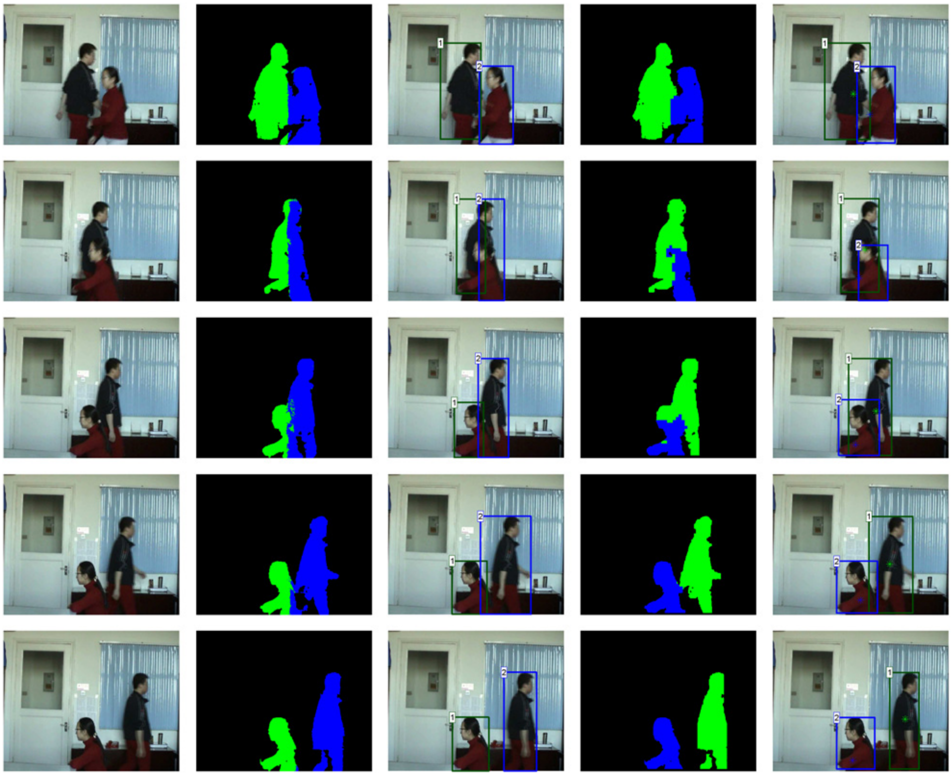

Fig. 14. Comparison results. Left column: The original image. The two columns in the middle: The results using the method of Elgammal and Davis [6]. The two columns in the right: The results using our new approach. (The frame numbers from the top to the bottom are respectively 34, 36, 38, 40, 42.)

They all show the tracking results of multiple objects through occlusion. Some data are got from the PETS datasets [27], while some others are typical data captured by ourselves. Out of these, four are outdoor data, while two are indoor data. Some of the data are gray images, whereas others are color images for comparison with other method. The labeling results of foreground regions are represented by points of different colors.

The tracking results of two persons' occlusion are shown in Fig. 9. With the similarity of the local texture between consecutive frames, our approach can track two persons successfully. The occlusion occurs between frame 364 and frame 385.

Experimental results show that our approach can track simultaneously different kinds (unknown) of objects. Fig. 10 illustrates key frames where a person and a vehicle are moving through severe occlusion. The data are obtained from PETS datasets [27]. The occlusion occurs between frame 552 and frame 587. Obviously, prior model based people tracking algorithm [6] may be unsuitable.

In some cases, the position changes of the moving objects between consecutive frames are irregular because of the velocities of objects are irregular or the frames are skipped. Thus the position prediction based tracking approach may cause error through occlusion. Experiments show that our approach can stably track objects with large frame skipping. The new approach can also track objects through occlusion even if the ground is invisible. Fig. 11 illustrates key frames where three persons are tracked through severe occlusion. The occlusion occurs between frame 67 and frame 112. The occlusion continues for about 3 s and the ground is invisible.

The new approach can stably track multiple objects with irregular movement after a long duration of complete occlusion. Fig. 12 shows the passing of two persons, where both of them at some instant are completely occluded by a tree. The first person paused for a while behind the tree, while the second person passed by the tree. After the second person has been visible again, the first person returned back. From frame 122 to frame 296, the first person is completely occluded by the tree. From frame 253 to frame 309, the second person is completed occluded by the tree.

Our new approach is proposed to segment and track multiple objects through occlusion by appearance based object modeling, and has the same input and output as in the work of Elgammal and Davis [6]. We make a comparison with the work of Elgammal and Davis [6], as shown in Figs. 13 and 14. Fig. 13 illustrates two persons with the similar blob density while the blobs are different in texture. When occlusion occurs, our approach can track well, but the method in Ref. [6] will fail to track. Fig. 14 illustrates that one person squats suddenly during occlusion. Even in this case, our approach works well, while the method in Ref. [6] makes a mistake. The degree of occlusion and the segmentation error rates of the two approaches are shown in Fig. 15.

Experimental results show that in contrast to the traditional approaches, especially Ref. [6], there are some advantageous features of our approach.

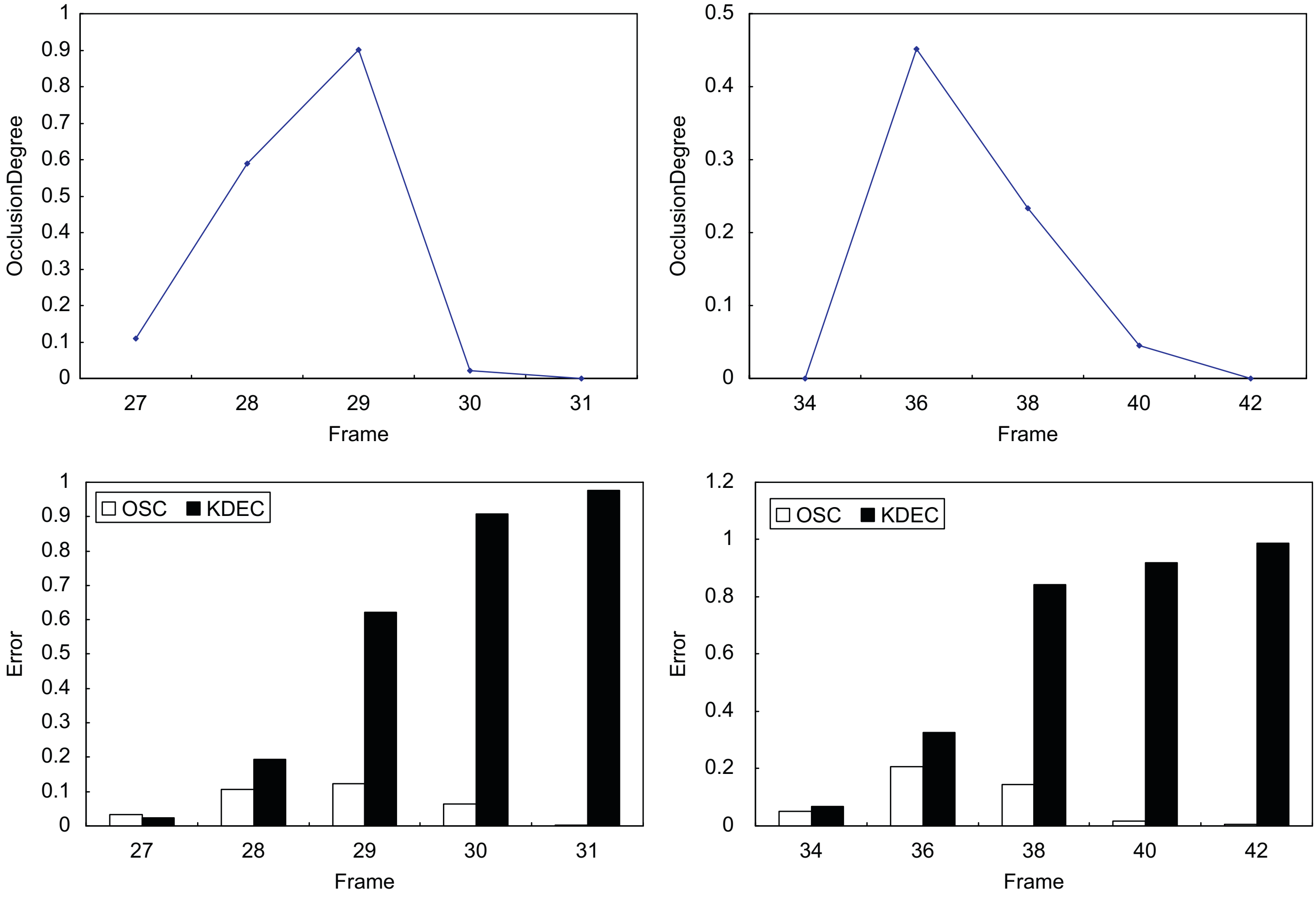

Fig. 15. Comparison results of the approach in this paper (OSC) and the approach of Elgammal and Davis [6] (KDEC). Top row: The occlusion degree of sequence in Figs. 13 and 14. Bottom row: The error of segmenting results, respectively.

Firstly, as shown in Fig. 10, our approach can stably track multiple objects with prior spatial model unknown. Our approach does not rely on the prior object's spatial information. In Ref. [6], the prior spatial model of tracking objects is one of prior constraints known before tracking.

Secondly, our approach is block sample based classification, which can distinguish the differences of local texture, whereas in Ref. [6], the method is blob density based classification, which ignores the differences of local texture within the blobs.

Thirdly, method in Ref. [6] assumes that the objects are all in upright pose, which is too restrictive for the real scene. Our approach does not rely on this assumption, and with block sampling, our approach can track robustly with pose changing greatly.

Fourthly, our approach does not rely on the assumption of only smooth changes in position between consecutive frames, and can well estimate the position of objects through occlusion in the case of irregular movement or frame skipping. Remarkably, the method in Ref. [6], however has an assumption of slight changes in position between consecutive frames, to reduce the large computational cost for position estimation, which may cause tracking error in complex cases.

Since our approach has relatively many parameters, naturally there arises a question: How easy or difficult is it to obtain a good set of parameter values? To see how sensitive the proposed method is to small changes of its parameter values, we calculated the error classifications in pixel for different parameter settings, where only one parameter was varied at a time. The measurements were made for several video sequences, including indoor and outdoor scenes. The results for the sequence of Fig. 9 are plotted in Fig. 16. It can be clearly seen that, for these parameters, a good value can be chosen across a wide range of values. The same observation was made for all the measured sequences. This property significantly eases the selection of parameter values. Furthermore, the experiments have shown that a good set of parameters for a sequence usually performs well also for other sequences.

The relationship of final label error to the error of center position estimation are shown in Fig. 17 for the sequence of Fig. 9. It shows that when the error of estimated center position to the real center position varies between −20 pixel and 20 pixel, the label error is within 15%. In our approach, the real error of estimated center position is within 5 pixel, which can slightly effect the final label error.

The method is implemented using C++ on a personal computer (with 2.0 GHz processor and 1 G memory). The image resolution is 320 × 240 pixels (24 bits per pixel). With the parameter values ($L = 15, \alpha = 0.5, n = 9, W = 30$), an average

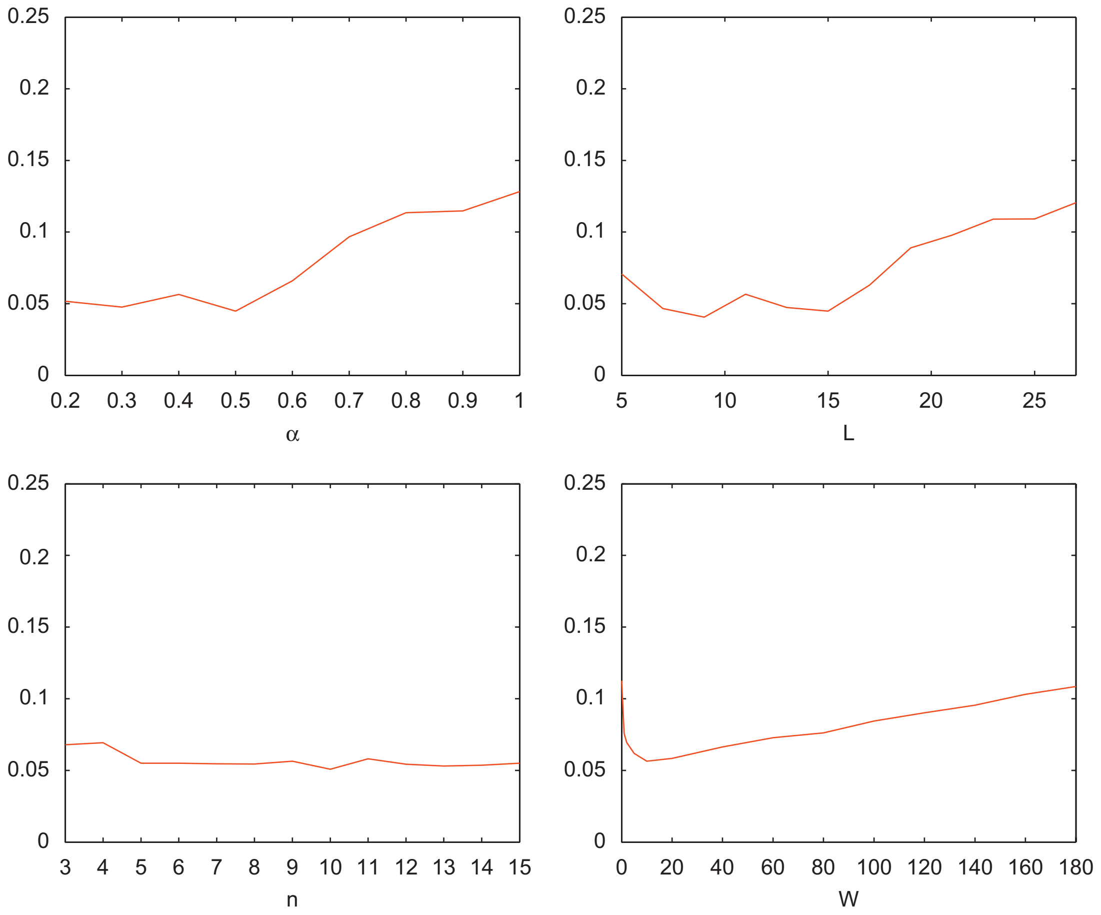

Fig. 16. Error rates for different parameters for the sequence of Fig. 9. While one parameter was varied, other parameters were kept fixed at the values: $\alpha = 0.5$, $L = 15$, $n = 9$, $W = 30$. The error rates are the total false negative rates of two objects averagely through occlusion.

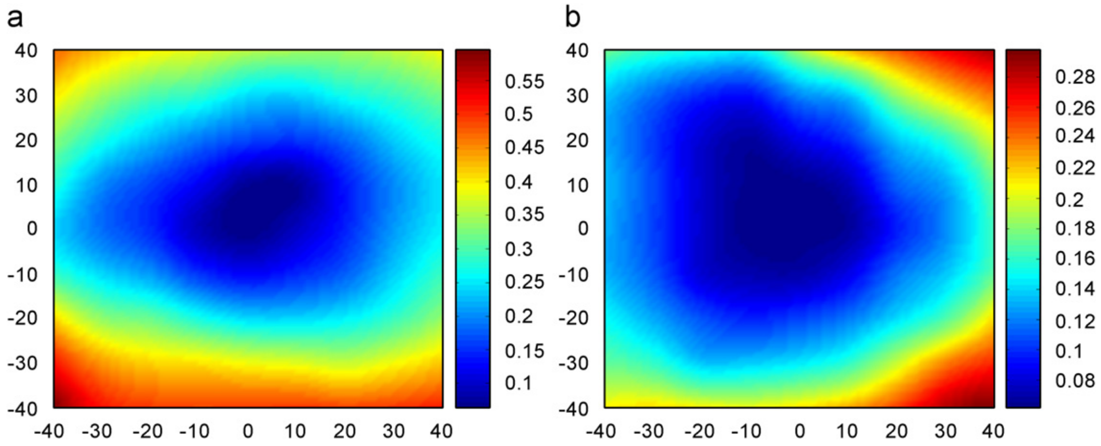

Fig. 17. The relationships between final label error and different center error for two objects for the sequence of Fig. 9. The 2D coordinates denote the error of estimated center position to the real center position. (a) Center error of object labeled-1 (Fig. 9) changes and center error of object labeled-2 is zero. (b) Center error of object labeled-2 changes and center error of object labeled-1 is zero.

processing rate of 11 frames per second is achieved by using our algorithm. When the objects are slightly overlapping, more regions need to be segmented. In this case, the processing rate of the proposed algorithm is 8 frames per second. It shows the proposed algorithm can meet the need of real-time applications.

4. Conclusion

In this paper, we introduce a novel online sample based classification approach to track multiple objects through occlusion. Training samples are extracted from the objects' foreground regions before occlusion. And two-stage classifiers are learned from these online updated training samples. When occlusion occurs, the foreground regions corresponding to multiple objects are extracted to several testing samples. Using the first-stage classifier with the local color and texture features to select spatial distinguishing points, the positions of the centers of multiple objects are estimated quickly. Then, using the second-stage classifier with the relative spatial features and the local color and texture features, the testing samples are labeled correctly. The contribution of our new approach can be summarized as follows.

Firstly, the center positions of multiple objects through occlusion are estimated accurately with unsmooth motion in consecutive frames, such as frame skipping, long-term severe occlusion and large pose changes.

Secondly, the instantaneous spatial features relative to the center of objects are used instead of the fixed spatial models. So, the new approach can track various kinds (unknown) of objects.

Thirdly, with novel online block sampling and classification with instantaneous local color and texture features and relative spatial features to the centers of the objects, the new approach can segment and track multiple objects through occlusion effectively.

Finally, the new approach can easily be combined with other methods to improve the accuracy of the tracking system in certain cases.

Future work includes improvement of feature description and combining with other methods to improve the accuracy of classification in real system.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 60573062 and 60673106) and the Specialized Research Fund for the Doctoral Program of Higher Education.

References

[1] A. Amer, E. Dubois, A. Mitiche, Real-time system for high-level video representation: application to video surveillance, in: Conference on Visual Communication and Image Processing (VCIP), Proceedings of the SPIE International Symposium on Electronic Imaging, vol. 5022, Santa Clara, USA, January 2003, pp. 530–541.

[2] R. Collins, A. Lipton, T. Kanada, et al., A system for video surveillance and monitoring: VSAM final report, Technical report CMU-RI-TR-00-12, Carnegie Mellon University, May 2000.

[3] I. Haritaoglu, D. Harwood, L.S. Davis, W4: real-time surveillance of people and their activities, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 809–830.

[4] F. Lv, J. Kang, R. Nevatia, I. Cohen, G. Medioni, Automatic tracking and labeling of human activities in a video sequence, in: Proceedings of the 6th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS04), Prague, Czech Republic, May 2004.

[5] C. Stauffer, W. Grimson, Learning patterns of activity using real time tracking, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 747–767.

[6] A.M. Elgammal, L.S. Davis, Probabilistic framework for segmenting people under occlusion, in: Proceedings of the 8th IEEE International Conference on Computer Vision, IEEE Computer Society, Los Alamitos, 2001, pp. 145–152.

[7] M. Hotter, R. Thoma, Image segmentation based on object oriented mapping parameter estimation, Signal Processing 15 (3) (1988) 315–334.

[8] K.P. Lim, A. Das, M.N. Chong, Estimation of occlusion and dense motion fields in a bidirectional Bayesian framework, IEEE Trans. Pattern Anal. Mach. Intell. 24 (5) (2002) 712–718.

[9] Q. Zang, R. Klette, Object classification and tracking in video surveillance, in: Proceedings Computer Analysis of Images and Patterns, Lecture Notes in Computer Science, vol. 2756, Springer, Berlin, 2003, pp. 198–205.

[10] T. Zhao, R. Nevatia, Tracking multiple humans in crowded environment, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, IEEE Computer Society, 2004, pp. 1I-406–II-413.

[11] T. Zhao, R. Nevatia, Tracking multiple humans in complex situations, IEEE Trans. Pattern Anal. Mach. Intell. 26 (9) (2004) 1208–1221.

[12] A. Senior, A. Hampapur, YL. Tian, L. Brown, S. Pankanti, R. Bolle, Appearance models for occlusion handling, in: Proceedings of the 2nd IEEE International Workshop on PETS, Kauai, Hawaii, USA, December 9, 2001.

[13] A. Senior, Tracking people with probabilistic appearance models, in: Proceedings of the ECCV Workshop on Performance Evaluation of Tracking and Surveillance Systems, June 1, 2002, pp. 48–55.

[14] H. Wang, D. Suter, Tracking multiple humans in complex situations, IEEE Int. Conf. Image Process. 2 (II) (2005) 410–413.

[15] L.-Q. Xu, P. Puig, A hybrid blob- and appearance-based framework for multi-object tracking through complex occlusions, in: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, October 15–16, 2005, pp. 73–80.

[16] M. Xu, T. Ellis, Partial observation versus blind tracking through occlusion, in: Proceedings of the BMVC, 2002, pp. 777–786.

[17] F. Fleuret, R. Lengagne, P. Fua, Fixed point probability field for complex occlusion handling, 10th IEEE International Conference on Computer Vision, vol. 1, 2005, pp. 694–700.

[18] C.C.C. Pang, W.W.L. Lam, N.H.C. Yung, A novel method for handling vehicle occlusion in visual traffic surveillance, in: Image Processing: Algorithms and Systems II, Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE), vol. 5014, SPIE-International Society of Optical Engineering, Bellingham, 2003, pp. 437–447.

[19] P. Guha, A. Mukerjee, K.S. Venkatesh, Efficient occlusion handling for multiple agent tracking by reasoning with surveillance event primitives, in: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 49–56.

[20] T. Yang, Q. Pan, J. Li, S.Z. Li, Real-time multiple objects tracking with occlusion handling in dynamic scenes, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005 (CVPR 2005), vol. 1, June 20–25, 2005, pp. 970–975.

[21] D. Conte, P. Foggia, J.-M. Jolion, M. Vento, A graph-based, multi-resolution algorithm for tracking objects in presence of occlusions, Pattern Recognition 39 (4) (2006) 562–572.

[22] A.D. Jepson, D.J. Fleet, T.F. El-Maraghi, Robust online appearance models for visual tracking, IEEE Trans. Pattern Anal. Mach. Intell. 25 (10) (2003) 1296–1311.

[23] H.T. Nguyen, A.W.M. Smeulders, Fast occluded object tracking by a robust appearance filter, IEEE Trans. Pattern Anal. Mach. Intell. 268 (2004) 1099–1104.

[24] C.R. Wren, A. Azarbayejani, T. Darrell, A.P. Pentland, Pfinder: real-time tracking of the human body, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 780–785.

[25] M. Hu, W.M. Hu, T.N. Tan, Tracking people through occlusions, in: Proceedings of the 17th International Conference on Pattern Recognition, vol. 2, IEEE Computer Society, Los Alamitos, 2004, pp. 724–727.

[26] R. Lumia, L. Shapiro, O. Zuniga, A new connected component algorithm for virtual memory computers, Computer Vision Graphics Image Process. 22 (2) (1983) 287–300.

[27] PETS'2001. ⟨http://visualsurveillance.org/PETS2001⟩.

About the Author—LIN ZHU received the B.S. degrees with best honor from the Automatic Control Department, Northwestern Polytechnical University, Xi'an, China, in 2001. She is now a Ph.D. student in the Department of Automation, Tsinghua University, Beijing, China. Her research interests include computer vision, pattern recognition and visual surveillance.

About the Author—JIE ZHOU (M01, SM04) received the B.S. and M.S. degrees both from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively. He received the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology (HUST), Wuhan, China, in 1995. From 1995 to 1997, he served as a postdoctoral fellow in the Department of Automation, Tsinghua University, Beijing, China. Currently, he is a professor and assistant chair in the Department of Automation, Tsinghua University. His research areas include computer vision, pattern recognition, information fusion and image processing. In recent years, he has authored more than 10 papers in international journals and more than 40 papers in international conferences. He is an associate editor for the International Journal of Robotics and Automation. He received the Best Doctoral Thesis Award from HUST, the First Class Science and Technology Progress Award from Ministry of Education, China and the Excellent Young Faculty Award from Tsinghua University in 1995, 1998 and 2003, respectively. He is a senior member of the IEEE.

About the Author—JINGYAN SONG is a professor in the Department of Automation at Tsinghua University, Beijing. He received his Ph.D. in systems engineering and engineering management at the Chinese University of Hong Kong in 1999. His current research interests are in machine learning, robotics, intelligent control and intelligent transportation system.