Scene Event Recognition Without Tracking[1)]

Shaogang Gong　Tao Xiang
(*Department of Computer Science, Queen Mary, University of London, London* E1 4NS, U. K.)
(E-mail: sgg@dcs. qmul. ac. uk)

Abstract We present a novel approach to behaviour recognition in visual surveillance under which scene events corresponding to object behaviours are modelled as groups of affiliated autonomous pixel-level events automatically detected using Pixel Change Histories (PCHs). The Expectation-Maximisation (EM) algorithm is employed to cluster these pixel-level events into semantically more meaningful blob-level scene events, with automatic model order selection using modified Minimum Description Length (MDL). The method is computationally efficient allowing for real-time performance. Experiments are presented to demonstrate the effectiveness of recognising these scene events without object trajectory matching.

Key words Activity and behaviour recognition, event recognition, event versus trajectory based representation

1 Problem statement

Understanding visual behaviour captured in CCTV footage is fundamental in visual surveillance. We consider that visual behaviours of objects are underpinned by scene events that are defined by groups of spatio-temporally affiliated autonomous pixel-level events[1]. By autonomous, we imply that both the number of these events and their whereabout in the scene are to be determined automatically bottom-up without top-down labelling using predefined hypotheses.

Over the past decade, numerous efforts have been made to model object behaviours[2~5]. Most of which heavily relied upon segmentation and tracking of objects in the scene[6~11]. This is due to the fact that visual behaviours have traditionally been modelled through matching the trajectories of objects observed in a scene, either statically as templates or dynamically as state machines. This process critically relies upon the accuracy and consistency of object segmentation and tracking which are often ill-posed in a typical surveillance scenario due to the presence of multiple objects, occlusion, drastic lighting change and discontinuous motion, all contributing to the fragmentation and inconsistent labelling of object trajectories.

More recently, several attempts have been made to circumvent the problems intrinsic to the trajectory based matching approach to behaviour recognition. Instead of computing trajectories through object tracking, these methods focus on discrete semantic event correlation based on localised pixel-level event detection through learning[1,12~16]. In particular, object grouping and segmentation were avoided. However, these purely pixel-level based approaches can be sensitive to noise due to the lack of modelling spatial correlations in the image space. They can also be computationally expensive due to the large number of events to be monitored simultaneously.

To address this problem, we present in this work a method for learning higher-level scene events given pixel-level autonomous events but crucially without the need for matching object trajectories. In Section 2, Pixel Change History (PCH) is introduced for pixel-level events detection. PCHs are computed as local intensity temporal histories of individual pixels. Significantly, they can be computed very efficiently in real-time compared to other techniques such as multi-scale temporal wavelets[15]. PCHs are combined with an adap-

1) Supported by the UK EPSRC and DTI under the Management of Information Programme
Received November 26, 2002; in revised form April 5, 2003
收稿日期　2002-11-26; 收修改稿日期　2003-04-05

tive mixture background model to form a representation for detecting and classifying pixel-level events. They also provide the basis for computing higher-level scene events with clearer semantics. In Section 3, blob-level scene events are computed using unsupervised clustering based on Expectation-Maximisation (EM) with automatic model order selection usinga modified Minimum Descriptive Length (MDL) criterion. Experiments are presented in Section 4 to demonstrate that semantically more meaningful scene events were recognised consistently without object trajectory matching. Conclusions are drawn in Section 5.

2 Detecting pixel-level autonomous events

Our aim here is to define a suitable multi-scale temporal representation that is capable of distinguishing at the pixel level temporal scene change of different durations. Due to the large number of pixel-level changes to be monitored in each image frame, the representation must also be inexpensive for real-time performance. Temporal wavelets were adopted for such a multi-scale analysis[15]. However, the computational cost for such multi-scale temporal wavelets at the pixel level is very expensive. Alternatively, Motion History Image (MHI) was introduced to detect visual changes by keeping a history of change which decays over time. It has been used to build holistic motion templates for the recognition of human movement[17] and moving object tracking[18]. An important advantage of MHI is that although it is a representation of the history of pixel-level changes, only one previous frame needs to be stored. It is also easy to implement with minimal extra computational cost. However, at each pixel, explicit information about its past is mostly lost in MHI when current change is updated to the model. This is because that a change occurring in the current frame will make the MHI 'jump' to its maximal value. To overcome this problem, Pixel Energy History was introduced to measure the mean magnitude of pixel-level temporal energy over a period of time defined by a backward window[13]. The size of the backward window determines the number of frames (history) to be stored. However, this approach suffers from sensitivity to noise and also being relatively expensive to compute.

2.1 Computing pixel change history (PCH)

Here we propose a new representation, Pixel Change History (PCH), for describing multi-scale temporal change at the pixel-level based on computing both the Motion History Image and Pixel Signal Energy. It is important to point out that this measurement is different from that computed by multi-scale spatio-temporal filtering widely adopted for estimating apparent image motion such as optic flow. No spatio-temporal correspondence is established. The PCH of a pixel is defined as:

$$P_{\zeta,\tau}(x,y,t)=\begin{cases}\min\left(P_{\zeta,\tau}(x,y,t-1)+\dfrac{255}{\zeta},255\right), & \text{if } D(x,y,t)=1\\ \max\left(P_{\zeta,\tau}(x,y,t-1)-\dfrac{255}{\tau},0\right), & \text{otherwise}\end{cases} \tag{1}$$

where $P_{\zeta,\tau}(x,y,t)$ is the PCH for a pixel at (x,y), $D(x,y,t)$ is a binary image indicating the foreground region, ζ is an accumulation factor and τ is a decay factor. When $D(x,y,t)=1$, instead of jumping to the maximum value, the value of a PCH increases gradually according to the accumulation factor. When no significant pixel-level visual change is observed in the current frame, pixel (x,y) will be treated as part of background and the corresponding pixel change history starts to decay. The speed of decay is controlled by the decay factor ζ. The accumulation factor and the decay factor give us the flexibility of characterising the pixel-level change over time. In particular, large values of ζ and τ imply that the history of visual change at (x,y) is considered over a longer backward temporal window. In the meantime, the ratio between ζ and τ determines how much weight is put on the recent change.

We consider that Motion History Image is a special case of PCHs in that a combined PCHs of all the pixels over a sequence of images is equivalent to the Motion History Image of the image sequence when ζ is set to 1. Furthermore, similar to that of Pixel Signal Energy[13], a PCH also captures a zero order pixel-level change, i. e. the mean magnitude of change over time. In addition, it is capable of capturing higher order temporal changes occurred at a pixel over time including speed, trend (uphill or downhill) and the phase of a change.

2.2 Pixel-level event detection

The significance of any localised pixel-level change depends on the underlying object activities and behaviours they are associated with. We ultimately wish to have a completely automated method to extract scene level semantics from local pixel-level visual change. We begin by considering the problem of detecting and differentiating pixel-level changes that are semantically significant at the scene level. For example, in a busy scene in the public place such as in a supermarket, we are interested in automatically detecting and classifying localised and persistent movement of objects (e. g. people stop and browse) and changes to the background (e. g. the introduction of new objects into the scene or the removal of existing objects from the scene). To this end, we wish to compute pixel-level events using both adaptive mixture background modelling and PCHs.

Adaptive mixture background models are commonly used to memorise and maintain the background color distribution of a dynamic scene[9,10,13]. The major strength of such a model is its insensitivity to persistent movements of background objects such as waving tree leaves. However, an adaptive mixture background model cannot differentiate, although may still be able to detect the presence of, pixel-level changes caused by different types of scene events with different significance. In general, pixel-level change can be

1) short term caused by constant moving objects such as the waving tree leaves,

2) median term caused by the introduction of novel dynamics (of moving object),

3) long term caused by either the introduction of novel static objects into the scene, or the removal of existing objects from the scene.

We consider that only median and long term changes are of semantical significance and refer them as pixel-level events.

If the binary image $D(x,y,t)$ in Equation (1) is given by the temporal difference between the current frame and the dynamic background maintained by an adaptive mixture model, then a PCH based foreground model can be introduced to not only detect the median and long term pixel-level changes but also filter out the short term changes associated with the background. More precisely, we delimitate pixel-level events as foreground pixels that satisfy:

$$P_{\zeta,\tau}(x,y,t) > T_H \tag{2}$$

where T_H is a threshold. We can further detect those events that are associated median term change if

$$| I(x,y,t) - I(x,y,t-1) | > T_M \tag{3}$$

where T_M is a threshold. Events that do not satisfy the above condition are caused by long term change such as the introduction of static novel objects into the scene or the removal of existing objects from the scene. For example, a pixel-level event caused by a browsing person and a pixel-level event caused by the removal of an object from a shelf in a shopping mall may have very similar PCH value, but the former event satisfies Condition (3) above while the latter does not, thus they are detected as different classes of events.

3 Recognising scene events

Recognition of scene level events for behaviour profiling has been attempted directly based on pixel-level events[13]. However, the large number of events detected and the noise sensitivity caused by ignoring spatial correlation among pixel-level events limit the success of such an approach. To address this problem, we consider unsupervised clustering (grouping) of pixel-level events not only according to spatial proximity but also by temporal correlation.

3.1 Grouping of pixel-level events

Let us first consider grouping pixel-level events spatially. The connected component method is adopted to group the detected pixel-level events into blobs, represented by bounding boxes. Small blobs are removed by a size filter. Those remaining blobs with an average PCH (of the PCHs for all the pixels within each blob) larger than a threshold T_B are considered as scene events. Each scene event is given by a 6-dimensional feature vector:

$$[x, y, w, h, R_f, R_m] \tag{4}$$

where (x,y) is the central position of the corresponding bounding box in the image, (w,h) is the bounding box dimension, R_f represents the percentage of the bounding box occupied by pixel-level events and R_m represents the percentage of those pixel-level events which satisfy Condition (3).

3.2 Scene event recognition using unsupervised clustering

Given detected scene events, we wish to automatically cluster them into different classes with corresponding semantics. This is performed by unsupervised clustering in the 6D feature space using Expectation-Maximisation (EM) with automatic model order selection using modified Minimum Description Length (MDL) principle[19,20].

Suppose there are n independent training data $\{y_1, \cdots, y_n\}$, belonging to class w and $w = \{1, \cdots, K\}$. The estimated model order $\hat{K}$ by a standard MDL algorithm is given by:

$$\hat{K} = \operatorname{argmin}\left\{-\sum_{i=1}^{n} \ln f(y_i \mid w, \hat{\theta}(K)) + \frac{\zeta(K)}{2}\ln(n)\right\} \tag{5}$$

where $f(y_i | w, \hat{\theta}(K))$ is the class-conditional density function, $\hat{\theta}(K)$ are the mixture parameters estimated by a maximum likelihood algorithm such as EM and $\zeta(K)$ is the number of parameters needed for a K-component mixture. If full covariance matrix is used, we have:

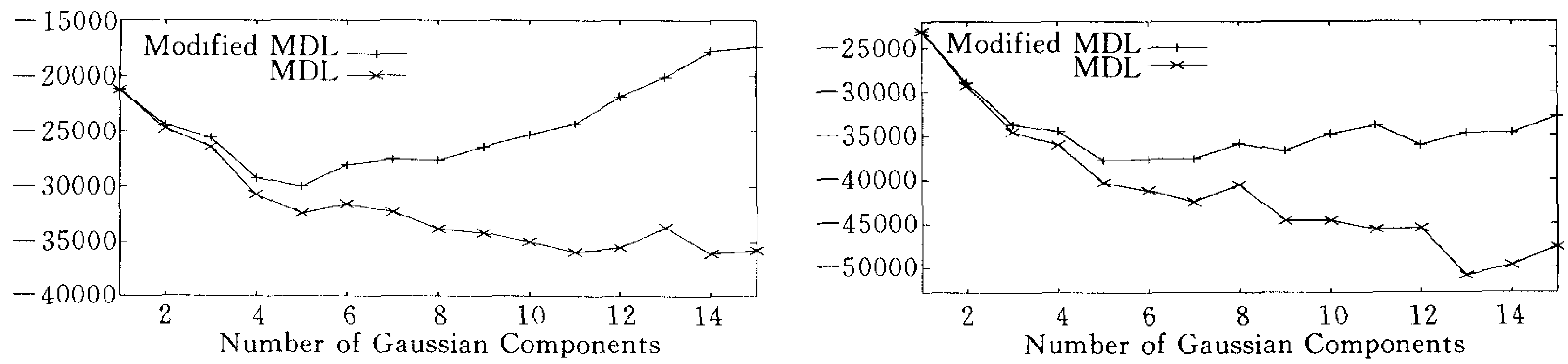
$$\zeta(K) = K - 1 + \frac{d^2 + 3d}{2}K \tag{6}$$

where d is the dimensionality of the feature space.

The first term in the bracket of Equation (5) corresponds to maximum likelihood, measuring the system entropy, while the second term measures the number of bits needed to encode the model parameters, serving as a penalty term for model complexity (i.e. very large K). One major problem with the standard MDL is that each component in the mixture can only 'see' the $m_j n$ subset of the data that has already been clustered to this component instead of the whole data set, where m_j is the weight for the j_{th} component. To overcome this problem, we adopt a modified MDL measure[20] with the model order $\hat{K}$ estimated as:

$$\hat{K} = \operatorname{argmin}\left\{-\sum_{i=1}^{n} \ln f(y_i / w, \hat{\theta}(K)) + \frac{K-1}{2}\ln(n) + \frac{d^2+3d}{4}K\ln(n)\right\} \tag{7}$$

The improvement from this modification over the standard MDL approach can be seen in Fig. 1. The resulting $\hat{K}$ gives the number of scene event classes and in a recognition



(a) Approach I

(b) Approach II

Fig. 1 Automatic model order selection using MDL and modified MDL. Model orders were considered in a range of (1,15)

process, a scene event is classified as one of the $\hat{K}$ with minimal Mahalanobis distance between the event and the mean of the cluster in the feature space.

4 Experiments

Experiments were conducted on a simulated 'shopping scenario' captured on a 20 minutes video at 25Hz. Some typical scenes and automatically detected pixel-level scene events are shown in Fig. 2. The 'scene' consists of a shop keeper sat behind a table on the right side of the view. Drink cans were laid out on a display table. Shoppers entered from the left and either browsed without paying or took a can and paid for it. An abnormal behaviour involves taking a can and leaving without paying. The data used for this experiment were sampled at 8 frames per second with total number of 5699 frames of images sized 320×240 pixels.

Two different approaches adopted for event detection are referred as Approach I and Approach II respectively as follows. For Approach I, only those foreground pixels that satisfy Condition (2) are detected as pixel-level events and all the blobs formed are recognised as scene events. For Approach II, all the foreground pixels are detected as pixel-level events and only those blobs with average Pixel Change History values larger than T_B are recognised as scene events. For the adaptive Gaussian mixture background model, the parameters were: learning rate $\alpha=0.002$, background model threshold $T=0.7$, six Gaussian components were maintained and a diagonal co-variance matrix was adopted. The parameters for pixel-level event detection were: $\zeta=12$, $\tau=10$, $T_H=180$, $T_M=10$ and $T_B=100$. Only those Blobs whose sizes were larger than 40 were considered. It was observed that using both approaches, localised movements such as "shopper paying" and the removal of background objects such as "can taken" were recognised automatically as scene events whist the occurrences of passing-by shoppers were ignored. For the whole 20 minutes scenario, 5019 and 4134 scene events were recognised using Approach I and Approach II respectively. Some examples of detected events are shown in Fig. 2. The algorithm was run on an Athelon 1.5G dual processor platform at an average speed of 6Hz without optimisation.

Unsupervised learning was performed on the first 3000 frames, where 2459 and 1922 scene events were detected using Approach I and Approach II respectively. EM was employed to obtain the parameters of the mixture model. It was combined with a modified MDL to determine the number of the classes of scene events and their whereabout. Both Approach I or Approach II automatically identified five different classes of scene events according to their location and temporal order through unsupervised clustering. They were labelled as "can taken", "entering and leaving", "shop keeper", "browsing" and "paying". For comparison, automatic model order selection using standard MDL is also shown in Fig. 1.

A testing set was composed using the rest of the frames from the 20 minutes video. The detected and classified autonomous events from this testing set were then projected onto the first three principal components of the 6D feature space (shown in Fig. 3). The spatial distributions of each class of events were illustrated by only showing their (x,y) co-ordinates of the central position of the corresponding bounding boxes in Figures 4,5,6, 7,8 and 9.

The model was used to perform online scene event recognition. Fig. 10 shows an example frame of the process. The extra computational cost was negligible and the algorithm still ran at 6Hz. For performance evaluation, the ground truth was labelled manually (see (a) and (b) in Fig. 11). Recognised events in each frame are shown in (c), (d), (e) and (f) of Fig. 11. To achieve a degree of robustness in event detection and classification, an event of a particular class was considered as presence if it has been recognised over a number of consecutive frames. Then, events were counted only once when they happened continuously. The detection rates and false detections of our algorithm in both individual classes of events and overall were measured against the ground truth and are shown in Table 1.

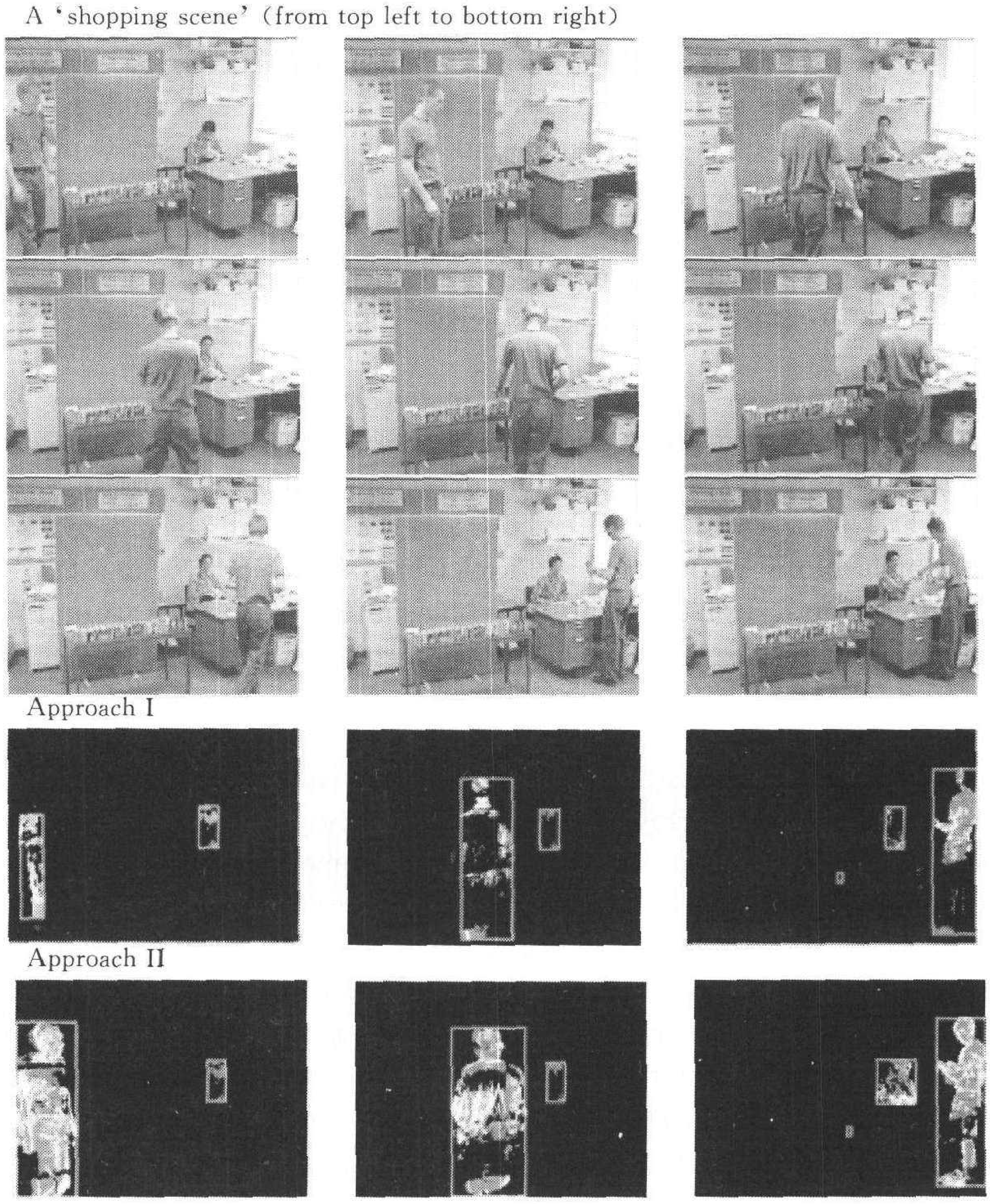
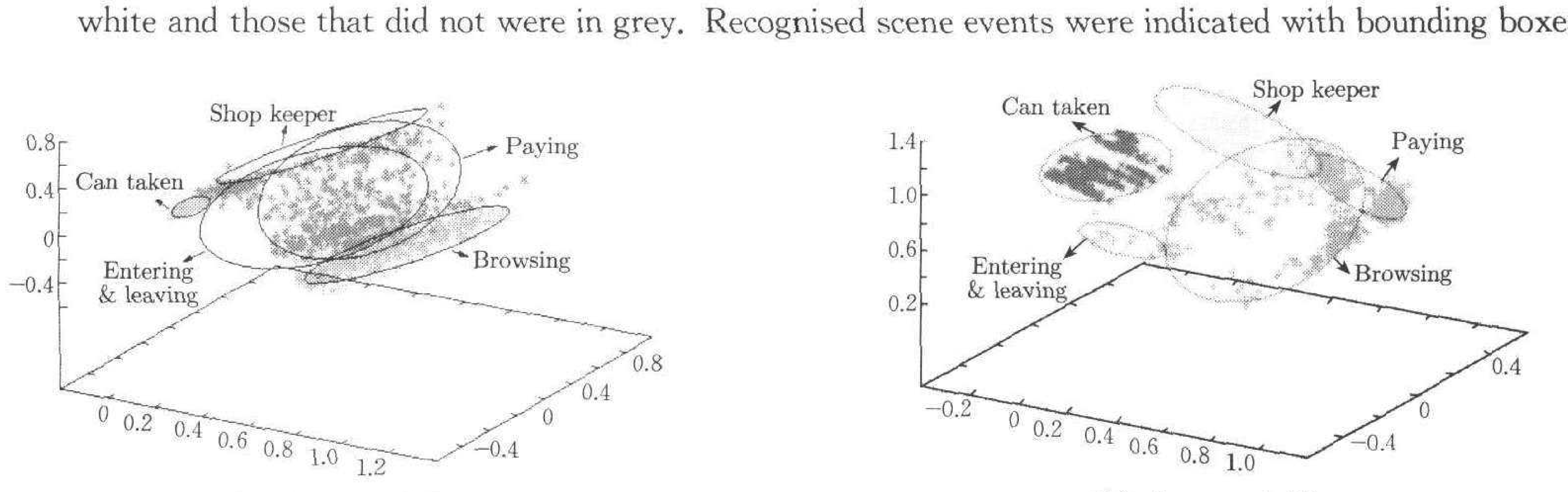


Fig. 2 Autonomous pixel-level event detection in a simulated shopping scenario. The figures in the top three rows from left to right, top to bottom are the typical scenes of the shopping scenario, which were sampled from frame 110 to frame 330 of the 20 minutes video. The figures in the fourth and the fifth rows are a number of events detected using Approach I and Approach II respectively. Pixel-level events that satisfied Condition (3) in Section 2.2 were highlighted in white and those that did not were in grey. Recognised scene events were indicated with bounding boxes



(a) Approach I (b) Approach II

Fig. 3 Event detection and classification of the testing set in the 6-dimeensional feature space (the first 3 principal components are shown)

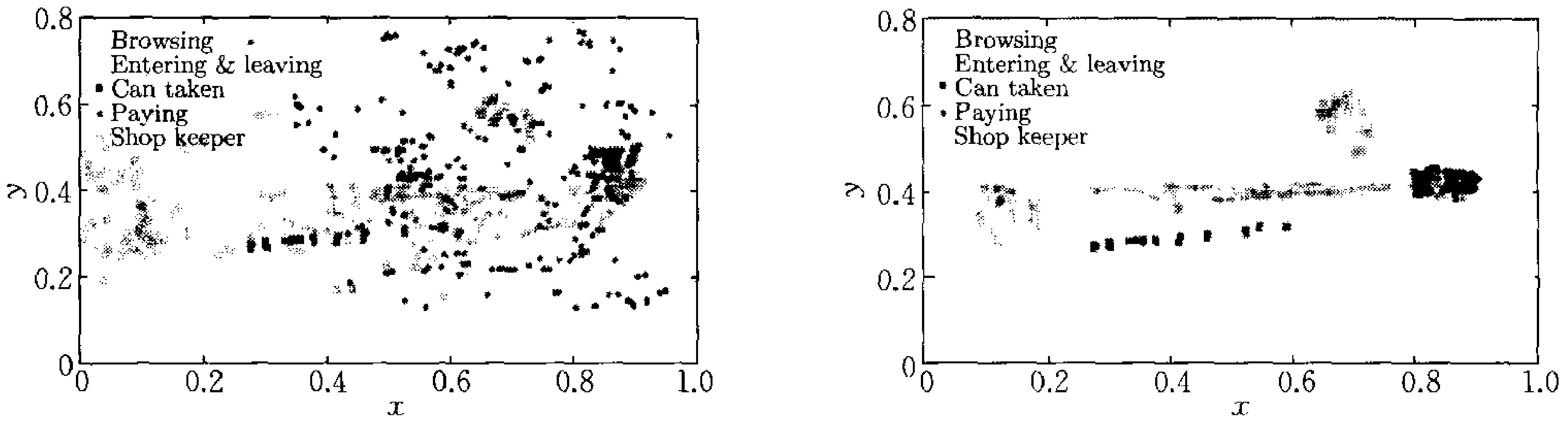


Fig. 4 Event detection and classification of the testing set in the image space. Top: 2560 scene events (not sustained) were detected using Approach-I, among which 929 were classified as"can taken" events, 283 as "entering and leaving" events, 293 as"shop keeper" events, 522 as "browsing" events and 533 as "paying" events. Bottom: 2212 scene events were detected using Approach-II, among which 1116 were classified as "can taken" events, 33 as "entering and leaving" events, 316 as "shop keeper" events, 406 as "browsing" events and 341 as "paying" events. Detected individual events of different classes are shown in Figures 6,7,8,9 and 10

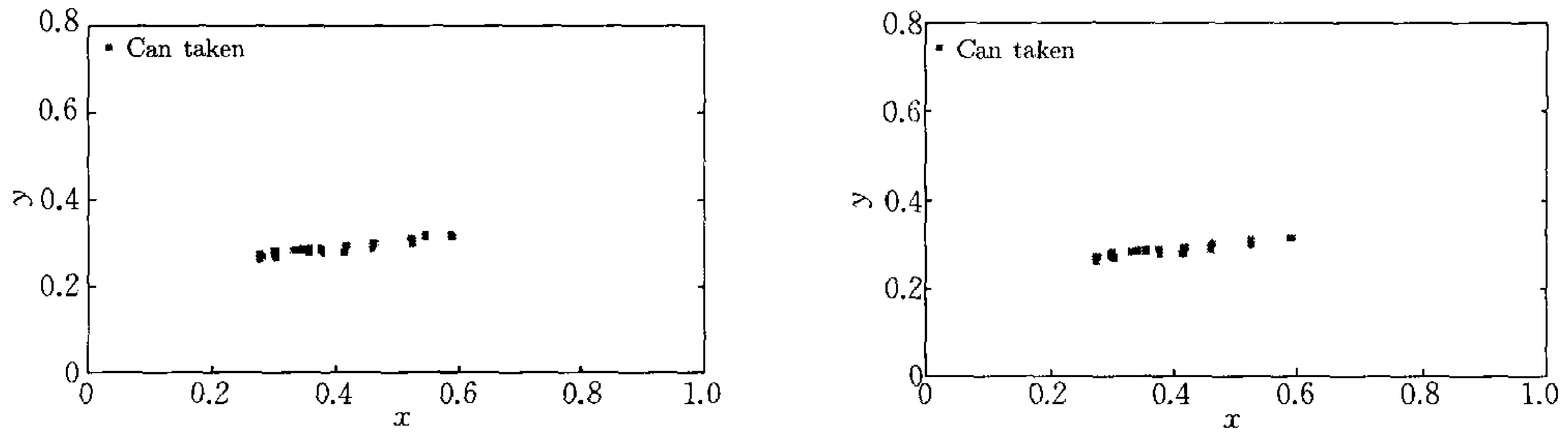


Fig. 5 Detected "can taken" events in the testing set. Top: 929 by Approach-I. Bottom: 1116 by Approach-II

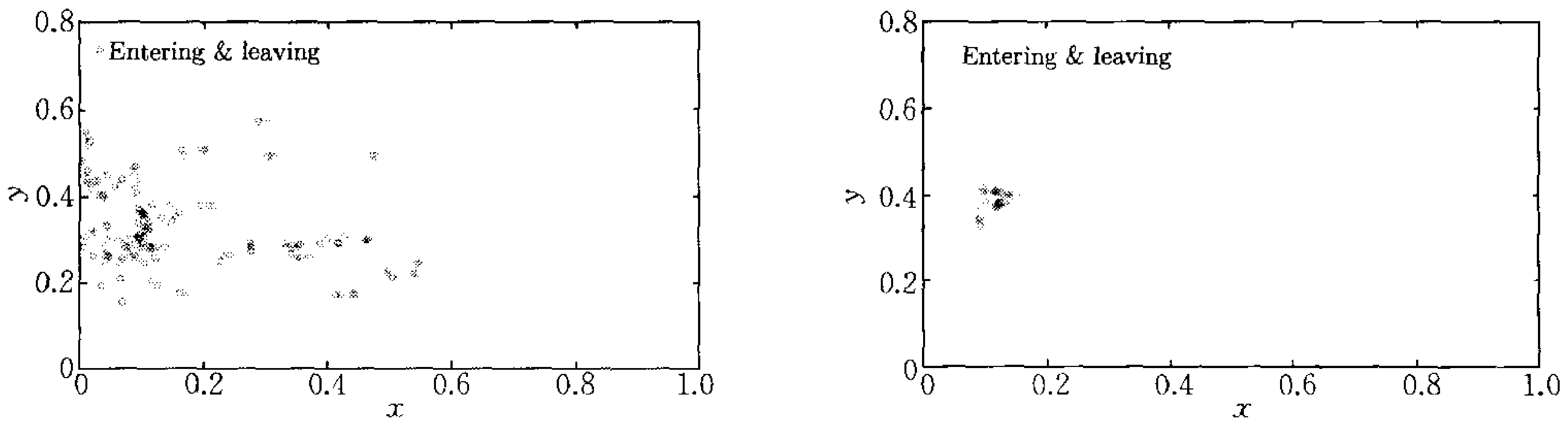


Fig. 6 Detected "entering and leaving" events in the testing set. Top: 283 by Approach-I. Bottom: 293 by Approach-II

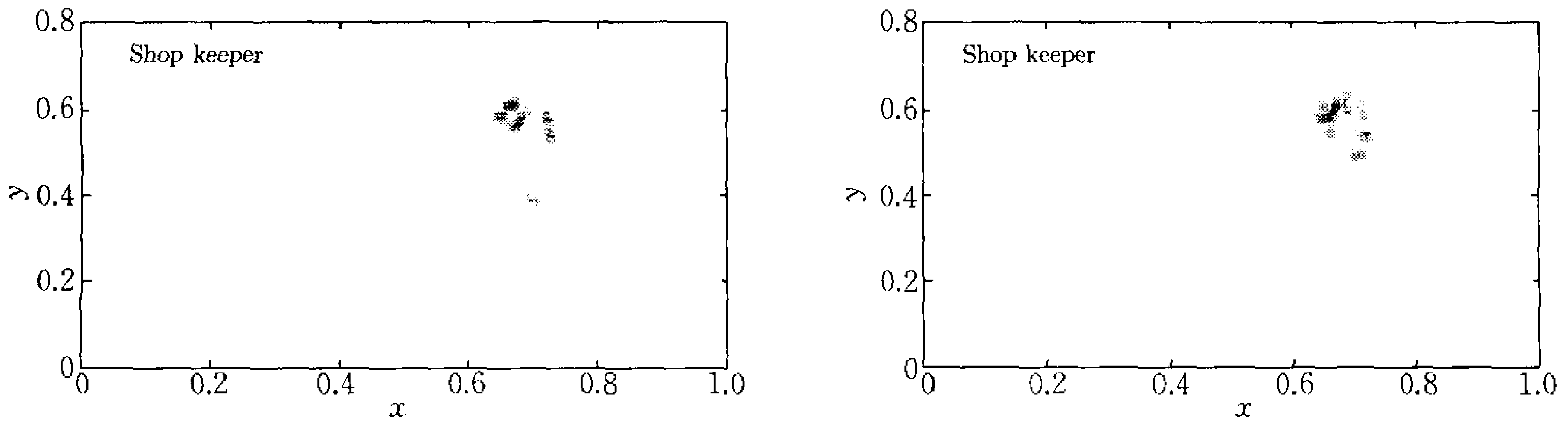


Fig. 7 Detected "shop keeper" events in the testing set. Top: 293 by Approach-I. Bottom: 316 by Approach-II

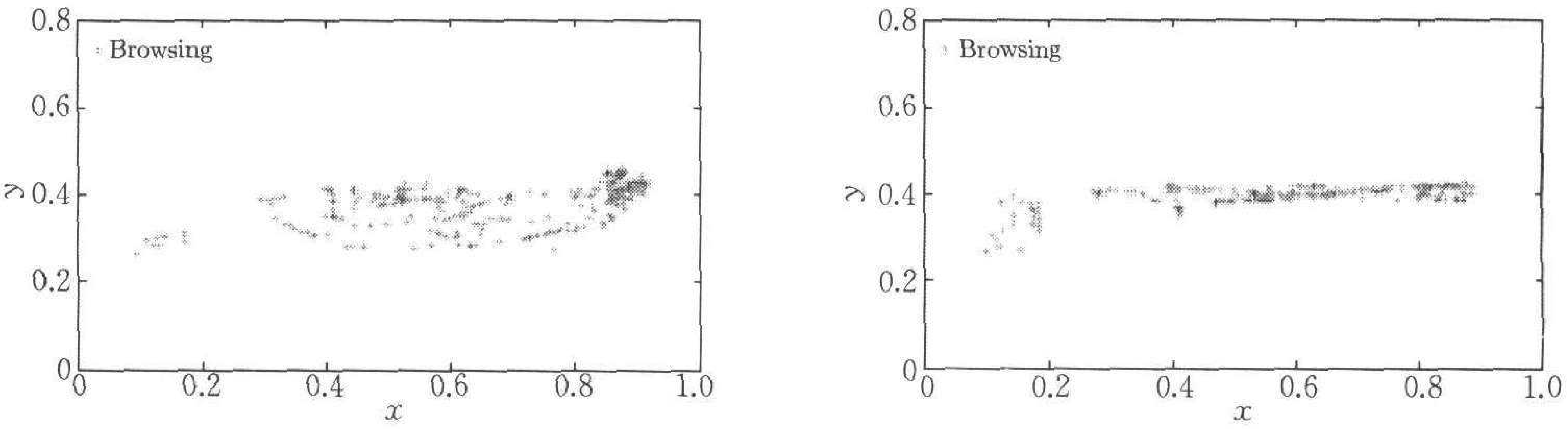


Fig. 8 Detected "browinig" events in the testing set. Top: 522 by Approach-I. Bottom: 406 by Approach-II

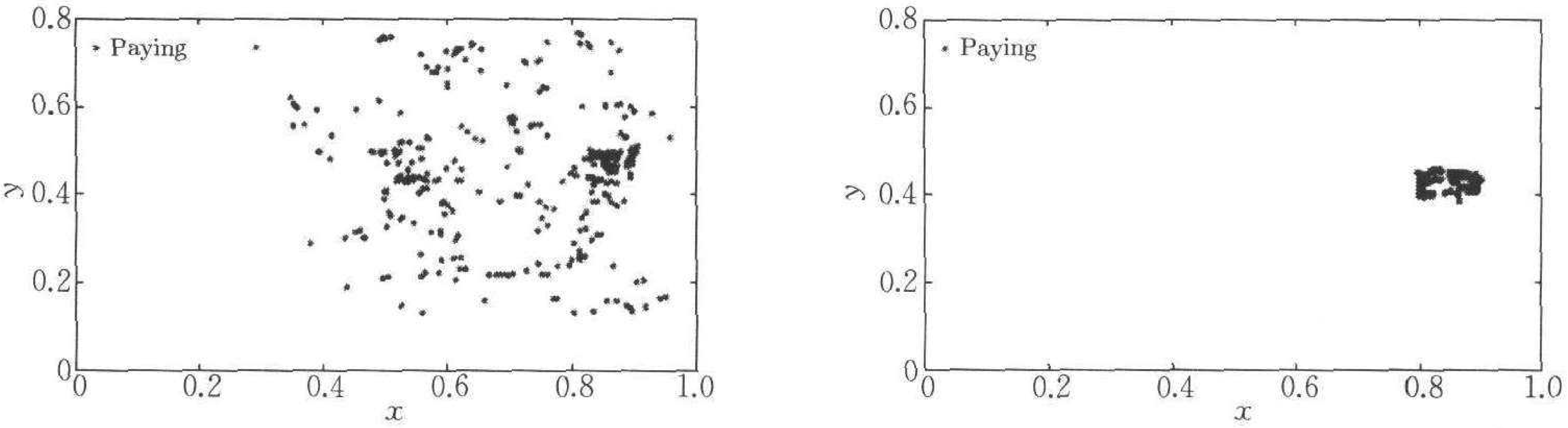


Fig. 9 Detected "paying" events in the testing set. Top: 533 by Approach-I. Bottom: 341 by Approach-II

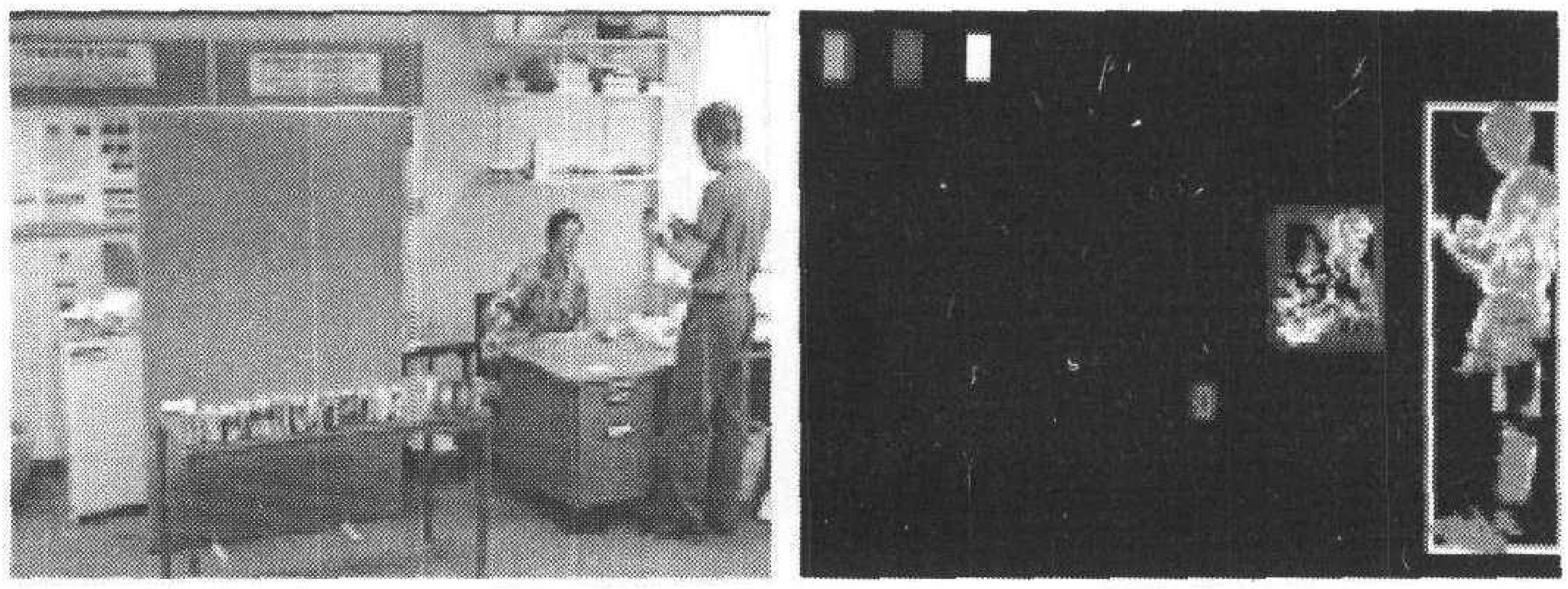
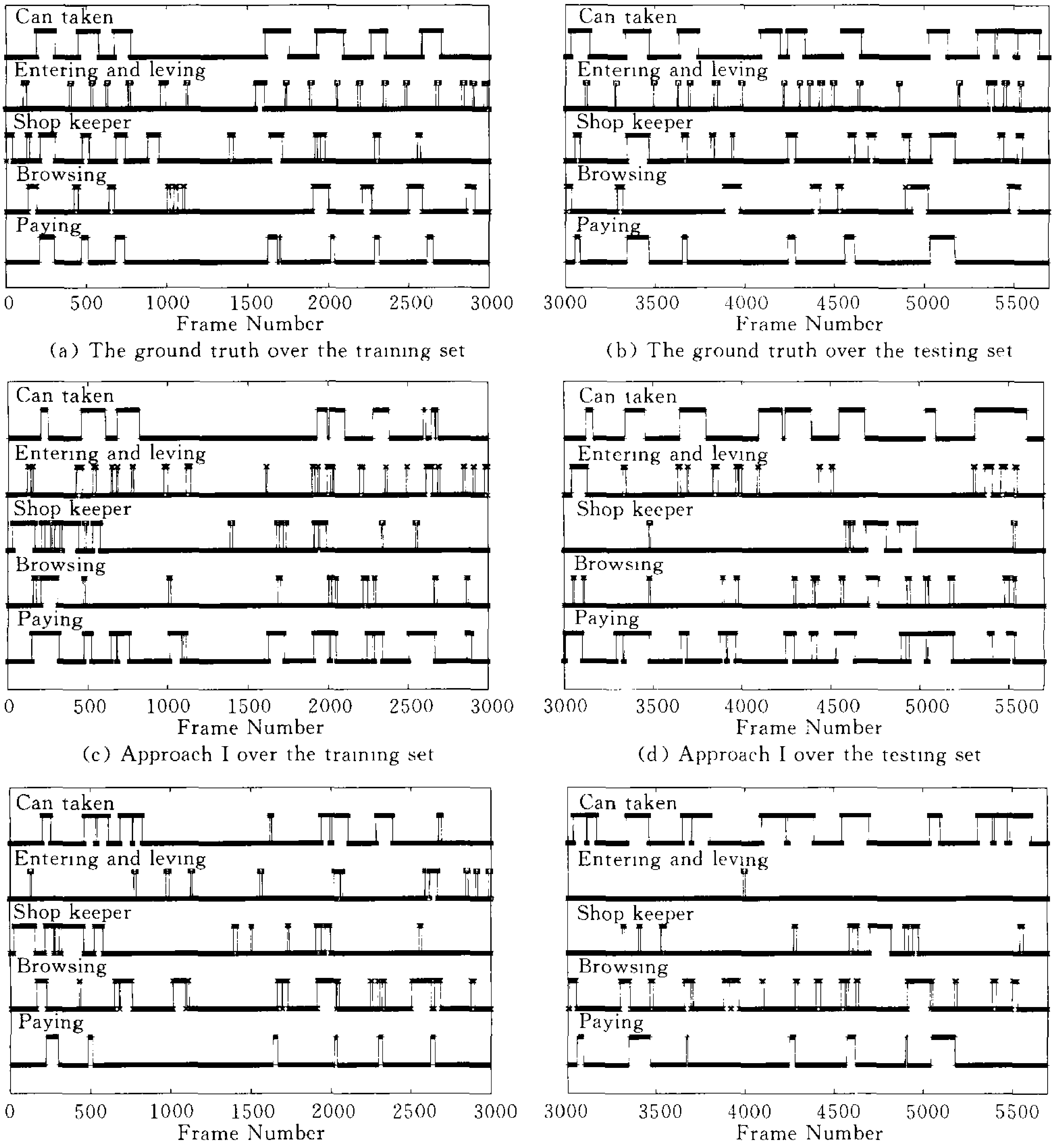


Fig. 10 Recognition of scene events at frame 3542 of the "can shop" sequence using Approach I. The left image shows the input frame and the right image depicts the output from the event recognition model. Events of "paying", "can taken" and "shop keeper" were indicated bounding boxes of different shades. Filled rectangular bins of the same shades were flashed out at the top-left corner of the screen to indicate the occurrences of events

Table 1 Event recognition rates and false detections. In the table, 'N' counts for the number of sustained events rather than the instantaneous events accumulated over time, i. e. a set of events of the same class detected continuously over successive frames was counted as 1 event of that class. 'App. I' and 'App. II' denote Approach I and Approach II respectively

Events	Training set					Testing set				
	N	Det. rate		False det.		N	Det. rate		False det.	
		App. I(%)	App. II(%)	App. I	App. II		App. I(%)	App. II(%)	App. I	App. II
Can taken	7	86	100	0	0	10	100	100	0	0
Ent. & lev.	18	67	56	8	1	18	61	6	3	0
Shop keeper	12	75	67	1	0	12	33	50	1	1
Browsing	10	60	100	3	7	8	63	100	9	10
Paying	8	100	75	6	0	6	100	100	6	1
Overall	55	75	75	18	8	54	67	57	19	12



(a) The ground truth over the training set

(b) The ground truth over the testing set

(c) Approach I over the training set

(d) Approach I over the testing set

(e) Approach II over the training set

(f) Approach II over the testing set

Fig. 11 Compare the ground truth with the recognised blob-level scene events. Each "can taken" event was counted for 100 frames in the ground truth

Results shown in Table 1 illustrate that scene events of "can taken" and "paying" were recognised accurately using both approaches, as was "browsing" using Approach II. The reason for the low recognition rate of "shop keeper" events was that the movements of the shop keeper were frequently occluded by the shoppers. Some shoppers entered and left the view without slowing down, thus no localised movement (median term change) was recognised in the scene, which resulted in the poor recognition rate of "entering and leaving". Other errors were mainly in the recognition of "paying" and "browsing" events. With Approach I, many "browsing" events were mistakenly recognised as "paying", leading to low recognition rate for "browsing" and large number of false recognitions for "paying". With Approach II, the starting and ending phases of "Paying", as well as some "entering and Leaving" events were frequently recognised as "browsing", leading to a large number of false recognitions of "browsing". A fusion of the two approaches could give more accurate recognition.

It was noticed that many "paying" and "browsing" events were spatially very close and featured similar movements. This will potentially pose a problem for the current model. For example, when a shopper stands in front of the shop keeper, it is impossible to tell whether he is going to pay or he is just browsing unless one takes into consideration whether any drink can was taken a moment ago. Even when the shopper has a can in hand, he still can walk back and continue browsing without paying. That is normal in any real shopping scenario. This suggests that one should not expect the system to resolve this ambiguity unless higher order spatio-temporal correlations among different classes of eventscan be fully explored. These correlations could be both spatial and temporal. The explicit modelling of such correlations among different classes of scene events provides the means for automatic extraction of high level semantics. This is our ongoing work.

5 Conclusion

To summarise, we present in this paper a novel approach to behaviour recognition in visual surveillance under which scene events and object behaviours are modelled as groups of affiliated autonomous pixel-level events automatically detected at the pixel-level using Pixel Change Histories (PCHs). PCH is employed as a more effective representation for modelling autonomous visual events at the pixel-level. The Expectation-Maximisation (EM) algorithm is employed to cluster these autonomous pixel-level events into semantically more meaningful blob-level scene events, with automatic model order selection using modified Minimum Description Length (MDL). The method is computationally efficient allowing for real-time performance. Our experiments show that such scene events can provide semantically meaningful interpretations without the need for object trajectory matching. The work done so far only represents the first step toward a more comprehensive model for behaviour recognition. Our future work will be focused on exploiting higher order spatio-temporal affiliations among different classes of events for automatic extraction of higher-level scene semantics.

References

1 Gong S, Ng J, Sherrah J. On the semantics of visual behaviour, structured events and trajectories of human action. *Image and Vision Computing*, 2002, **20**(12): 873～888

2 Aggarwal J K, Cai Q. Human motion analysis: A review. *Computer Vision and Image Understanding*, 1999, **73**(3): 428～440

3 Gavrila D M. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 1999, **73**(1): 82～98

4 Buxton H, Gong S. Visual surveillance in a dynamic and uncertain world. *Artificial Intelligence*, 1995, **78**(3): 431～459

5 Moeslund T, Granum E. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 2001, **81**(3): 231～268

6 Gong S, Buxton H. On the visual expectations of moving objects: A probabilistic approac with augmented hidden Markov models. In: Proceedings of European Conference on Artificial Intelligence, Austria: Vienna, 1992. 781～786

7 Haritaoglu I, Harwood D, Davis L S. W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, **22**(8): 809～830

8 Intille S, Davis J, Bobick A. Real-time closed-world tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, 1997. 697～703

9 McKenna S, Jabri S, Duric Z, Rosenfeld A, Wechsler H. Tracking group of people. *Computer Vision and Image Understanding*, 2000, **80**(1): 42～56

10 Stauffer C, Grimson W. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, **22**(8): 747～758

11 Wada T, Matsuyama T. Multiobject behavior recognition by event driven selective attention method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, **22**(8): 873～887

12 Sherrah J, Gong S. Continuous global evidence-based Bayesian modality fusion for simultaneous tracking of multiple objects. In: IEEE International Conference on Computer Vision, 2001. 42～49

13 Sherrah J, Gong S. Tracking discontinuous motion using Bayesian inference. In: European Conference on Computer Vision, 2000. 150～166

14 Ng J, Gong S. Learning pixel-wise signal energy for understanding semantics. In: British Machine Vision Conference, 2001. 695～704

15 Sherrah J, Gong S. Automated detection of localised visual events over varying temporal scales. In: European Workshop on Advanced Video-based Surveillance System, 2001

16 Chomat O, Martin J, Crowley J. A probabilistic sensor for the perception and the recognition of activities. In: Proceedings of European Conference on Computer Vision, 2000. 487～503

17 Bobick A, Davis J. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, **23**(3): 257～267

18 Piater J H, Crowley J L. Multi-modal tracking of interacting targets using Gaussian approximation. In: Proceedings of the 2nd IEEE Workshop on Performance Evaluation of Tracking and Surveillance, 2001. 141～147

19 Figueiredo M, Jain A K. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, **24**(3): 381～396

20 Vailaya A, Figueiredo M, Jain A K, Zhang H J. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 2001, **10**(1): 117～130

Shaogang Gong Received the bachelor degree (information theory & measurement) from the University of Electronic Sciences and Technology of P. R. China in 1985 and the Ph. D. degree (computer vision) from the University of Oxford in 1989. He was a recipient of a Sino-Anglo Queen's Research Scientist Award in 1987, a Royal Society Research Fellow in 1987 and 1988, a GEC sponsored Oxford industrial fellow in 1989, a Postdoctoral Research Fellow on the EU ESPRIT-II project VIEWS in 1989～1993. He joined the faculty of Department of Computer Science at Queen Mary College, University of London as a Lecturer in 1993, was made a Reader in 1999 and appointed as Professor of Visual Computation in 2001. His research interests include computer vision, visual synthesis and machine learning including dynamic scene understanding, motion-based recognition, generative dynamical models, face and gesture recognition, activity and behaviour recognition, expression and gesture synthesis, visually mediated interaction, visual surveillance, statistical learning and kernel methods, probabilistic graph models, Bayesian networks and hidden Markov models.

Tao Xiang Received the bachelor degree in electrical engineering from Xi'an Jiaotong University of P. R. China in 1995 and the Ph. D. degree in electrical and computer engineering from National University of Singapore in 2002. In 2001, he joined the Computer Science Department, Queen Mary College, University of London, as a Postdoctoral Research Fellow, where he is currently working on group activity modelling for visual surveillance. His research interests include computer vision, pattern recognition, and data mining.

无需跟踪的场景事件识别

Shaogang Gong Tao Xiang

(*Department of Computer Science, Queen Mary, University of London, London E1 4NS*, 英国)

(E-mail: sgg@dcs. qmul. ac. uk)

摘　要　提出了一种用于视觉监控中行为识别的新颖方法. 该方法将相应于目标行为的场景事件建模为一组使用 PCH(Pixel Change Histories)检测的自治像素级事件. 结合基于改进的 MDL (Minimum Description Length)的自动模型规则选择, EM(Expectation-Maximisation)算法被采用来聚类这些像素级的自治事件成为语义上更有意义的区域级的场景事件. 该方法是计算上有效的, 实验结果验证了它在不需匹配目标轨迹的情况下自动识别场景事件的有效性.

关键词　行为识别, 事件识别, 基于轨迹表达的相对事件

中图分类号　TP391.41