Recognition of Scene Events Without Tracking

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 and object behaviours are modelled as groups of affiliated autonomous events automatically detected at the pixel-level using Pixel Change Histories (PCHs). The Expectation-Maximisation (EM) algorithm is employed to cluster these pixel-level autonomous 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 with ease. Experiments are presented to demonstrate the effectiveness of recognising such scene level events (and behaviours) automatically without matching object trajectories and manual labelling.

Key words Autonomous event detection, behaviour profiling, expectation maximisation, modified minimum descriptive length, motion history image, pixel change history, scene events, trajectory matching

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 which in turn are defined by groups of spatio-temporally affiliated pixel-level autonomous events[1]. By autonomous events, we imply that both the number of these groups of autonomous events and their whereabout in the scene should be automatically detected bottom-up without top-down manual labelling using predefined hypotheses, as adopted by most of the existing techniques.

Over the past decade, numerous efforts have been made to model object behaviours[2~5]. Most of which heavily rely 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. The process relies critically on the accuracy and consistency of object segmentation and tracking which are unfortunately 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 associated with the trajectory matching based approach for behaviour recognition. They include semantical events correlation[12,13] and learning localised pixel-level scene change. In particular, object grouping and segmentation were avoided by either profiling behaviour based on autonomous pixel-level events[14,15] or extracting features for the whole image based on pixel-level analysis[16]. However, these purely pixel-level based approaches can be sensitive to noise due to the lack of modelling spatial correlations among neighbouring pixels. 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

Received November 26, 2002; in revised form March 23, 2003
收稿日期 2002-11-26; 收修改稿日期 2003-03-23

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 the local intensity temporal histories of individual pixels. Significantly, it can be computed very efficiently in real-time compared to other techniques such as multi-scale temporal wavelets[15]. PCHs are combined with an adaptive 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 using a modified Minimum Descriptive Length (MDL) criterion. Experiments are presented in Section 4 to demonstrate that semantically more meaningful scene events can be recognised consistently without matching object trajectories. 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 computationally 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 additional computational cost. However, at each pixel, explicit information about its past is mostly lost when current change is updated to the model since 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[14]. The size of the backward window determines the number of frames (history) needed to be stored. However, this approach suffers from sensitivity to noise and also being computationally expensive.

2.1 Computing pixel change history (PCH)

Here we propose a new representation, referred to as the Pixel Change History (PCH), for multi-scale temporal pixel-level change detection based on the principles of both 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 when computing a PCH for a pixel over time. More precisely, the PCH of a pixel is computed 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 through 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 the image frames 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[14], a PCH also captures a zero order pixel-level change, i. e. the mean magnitude of change over time. In addition, however, 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 events detection

The interpretation of scene level events (their semantics) that are associated with meaningful object activities and behaviours largely depend on the context of the scene. 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 caused by scene events of significantly different semantics. 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, let us first introduce the notion of pixel-level events and consider the roles of adaptive Gaussian mixture background models and by computing PCHs.

Adaptive mixture background models are commonly used to memorise and maintain the background color distribution[9,10,14]. 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 classes of scene events with significantly different semantics. Pixel-level change can either be short term caused by 1) constant moving objects such as the waving tree leaves, or median term caused by 2) the introduction of novel dynamics (of moving object) or long term caused by 3) the introduction of novel static objects into the scene, or 4) 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 Eq. (1) above 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 dynamic 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 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 changes 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[14]. However, the large number of events detected and the noise sensitivity caused by ignoring spatial correlation of 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 autonomous 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. If pixel-level events refer to all the foreground pixels, only those blobs with an average PCH (of the PCHs for all the pixels within each blob) larger than a threshold T_B will be considered as blob-level scene events and kept for further processing. Each blob-level scene event is given by a 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 events recognition using unsupervised clustering

After blob-level (scene) events are recognised, behaviour profiling can be performed by first clustering the events into different classes. Each class of blob-level event corresponds to a significant phase (such as the starting or ending) of a higher-level activity. In order to detect the presence of any meaningful events and their whereabout in the scene, clustering are performed in a 6-D feature space given by the feature vector defined in (4). Examples of this 6-D feature space are illustrated using the projection of the three largest principal components shown in Fig. 3. Depending on the representation of events, different unsupervised clustering methods can be employed. We adopt Expectation-Maximisation (EM) with automatic model order selection using modified Minimum Description Length (MDL) principle[19,20].

MDL is employed to extend maximum likelihood estimation to the model order unknown situation. Let us consider 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 astandard 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 Eq. (5) corresponds to the maximised 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 very complex mixtures (i. e. very large K). One major problem with the standard MDL lies on the fact that each component in the mixture can only 'see' the $m_j n$ data (m_j is the weight for the j_{th} component) belonging to it, instead of the whole dataset. We adopt a modified MDL measure[19] with the model order $\hat{K}$ estimated as:

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

The obtained parameters of the mixture model are used to classify blob-level scene events. More specifically, each correspondent feature point is classified into a class so that the Mahalanobis distance between the feature point and the mean of the class cluster is minimal.

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 and scene events are shown in Figure 1. 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.

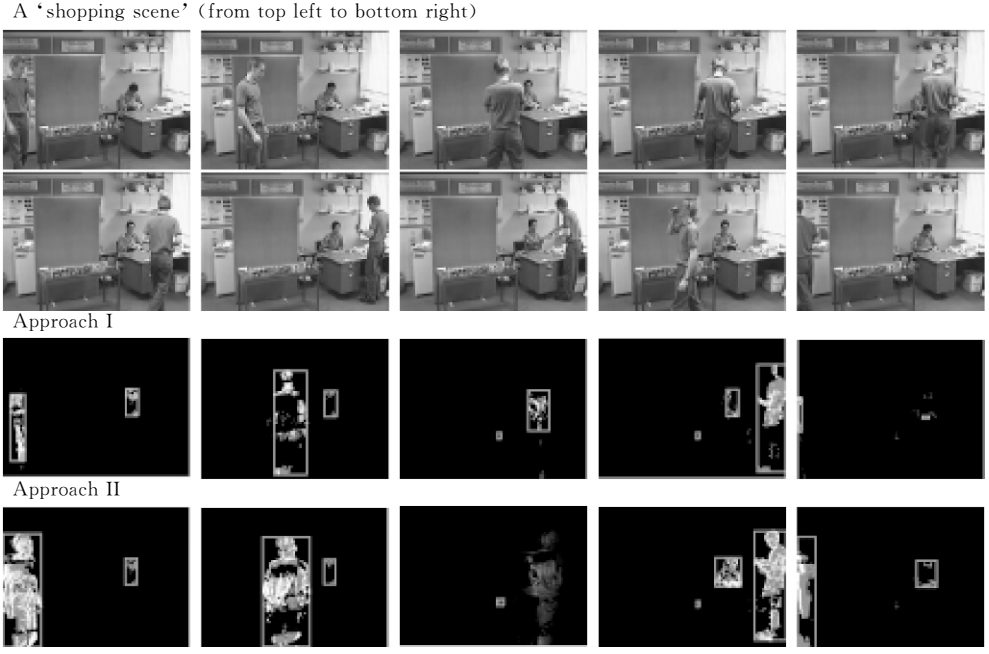


Fig. 1 Autonomous event detection in a simulated shopping scenario. The figures in the top two rows from left to right, top to bottom are the typical scenes of the shopping scenario, which were sampled from frame 110 to frame 330 of the 20 minutes video. The figures in the third and the fourth rows are a number of autonomous 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 blob-level scene events were indicated with bounding boxes

Two different approaches adopted for autonomous events 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 blob-level 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 blob-level scene events. For the adaptive Gaussian mixture background model, the parameters were set as: 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 events detection were chosen as $\zeta=12$, $\tau=10$, $T_H=180$, $T_M=10$ 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 significant events from visual changes, whist the occurrences of passing-by shoppers were ignored. For the whole 20 minutes scenario, 5019 and 4134 blob-level scene events were recognised using Approach I and Approach II respectively. Some of the results from this events detection process are shown in Fig. 1. 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 events and 1922 events were recognised 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 significant events in the scene and their whereabout. Fig. 2 shows that 5 classes of events were automatically recognised unsupervised using either Approach I or Approach II. For comparison, automatic model order selection using standard MDL is also shown in Fig. 2. Five different event classes were automatically learned in terms of their location and temporal order through unsupervised clustering, but with manual labelling to "can taken", "entering and leaving", "shop keeper", "browsing" and "paying".

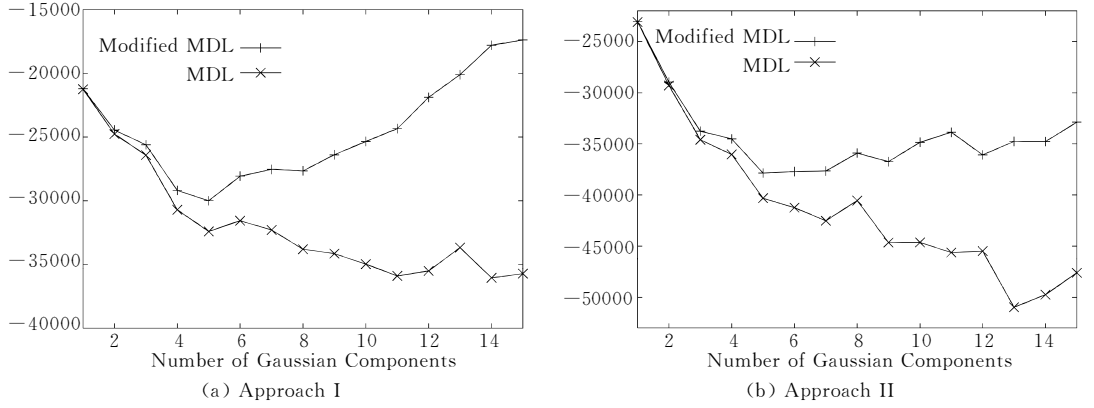
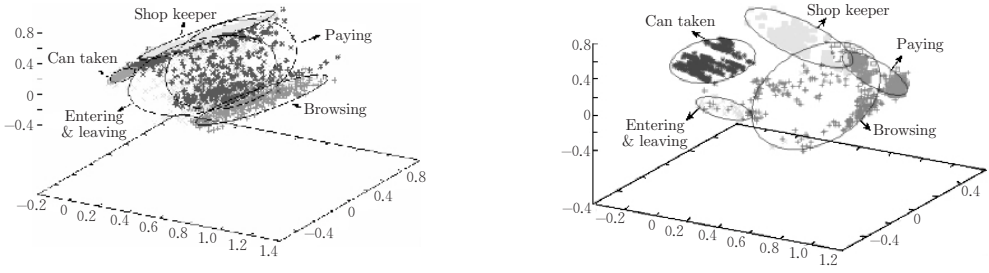


Fig. 2 Automatic model order selection using MDL and modified MDL. Model orders were considered in a range of (1,15)

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 three largest principal components of the 6-D 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 Figs. 4 and 5.



(a) Approach I (b) Approach II

Fig. 3 Autonomous events detection and classification on the testing set

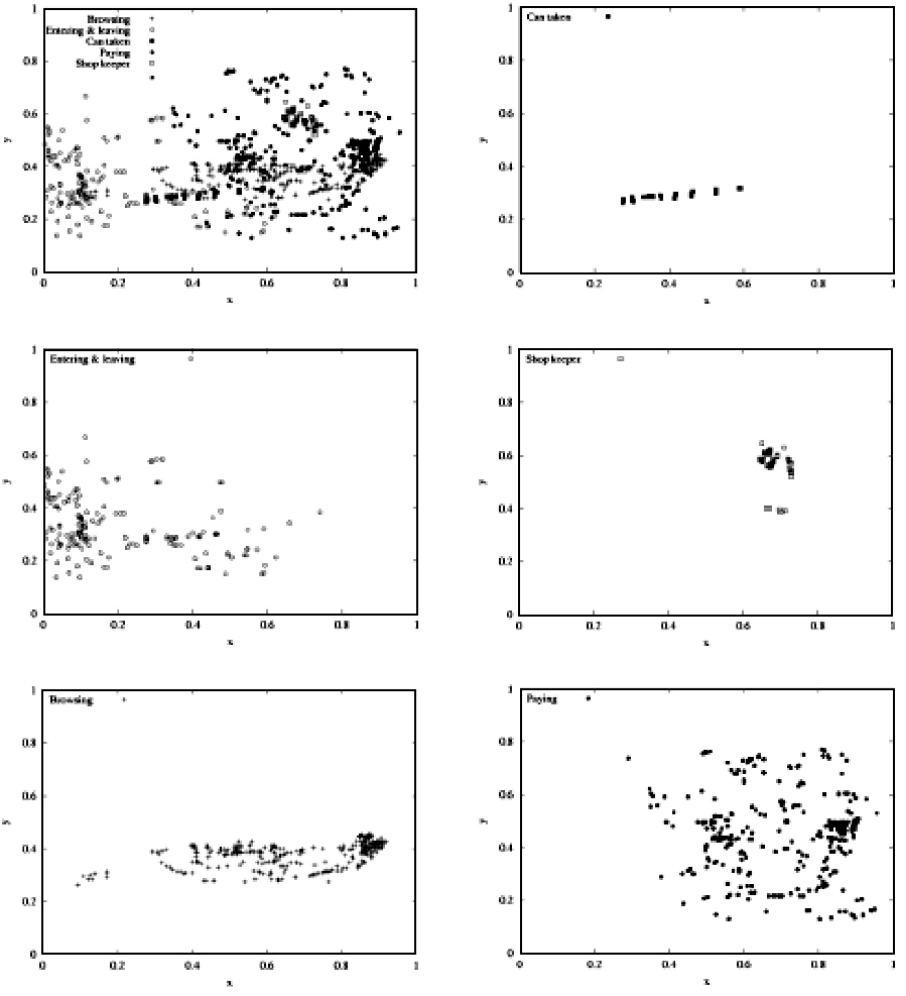


Fig. 4 Classification of the testing set in the image space using Approach I. From left to right, top to bottom, the figures are: 2560 blob-level scene events recognised from the testing set, (among which) 929 "can taken" events, 283 "entering and leaving" events, 293 "shop keeper" events, 522 "browsing" events and 533 "paying" events

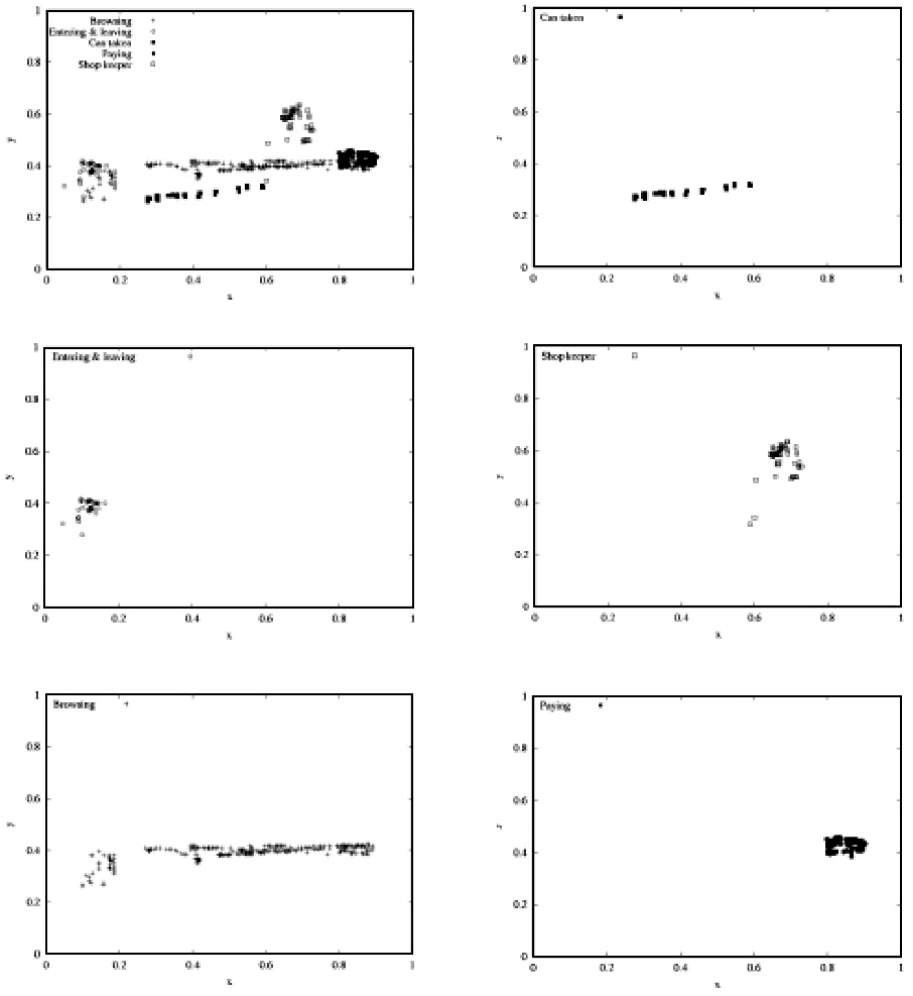
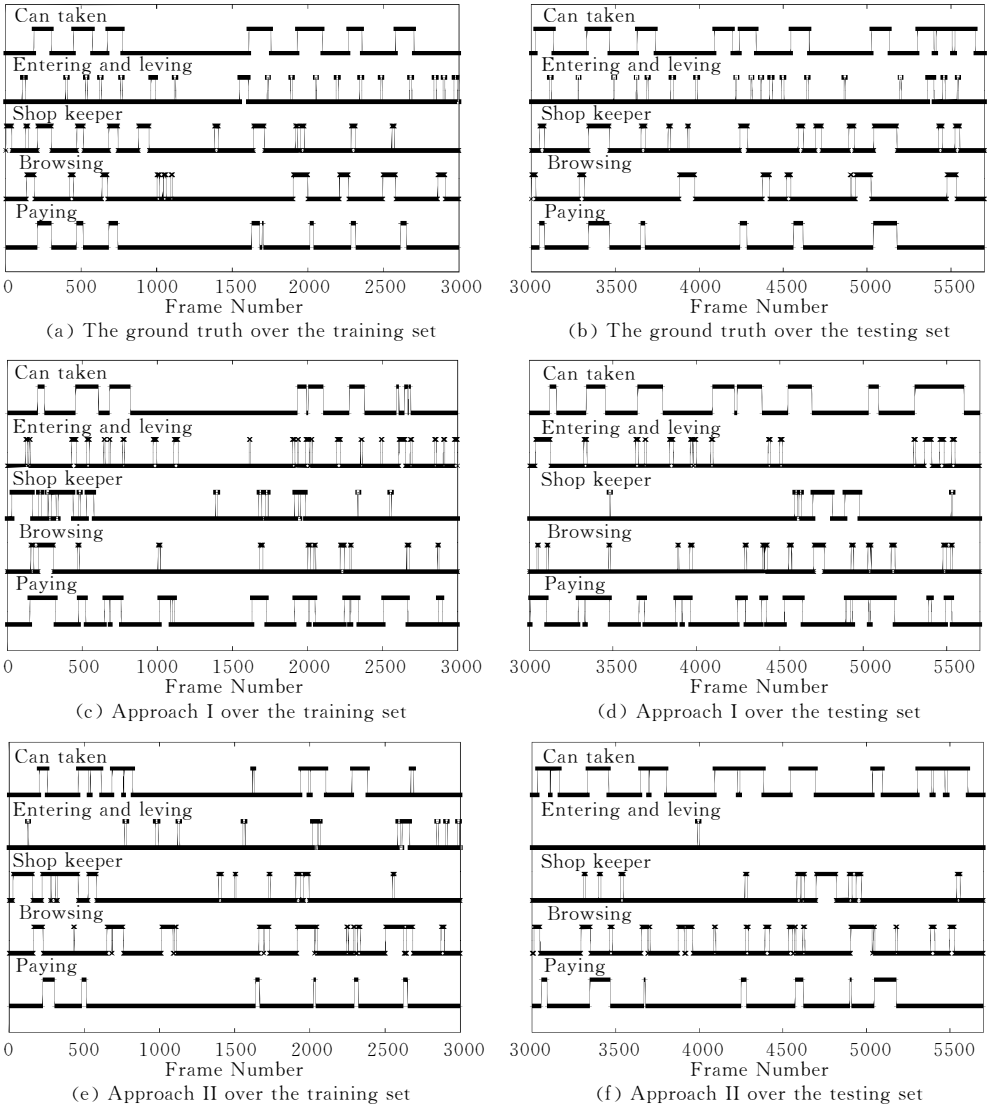


Fig. 5 Classification of the testing set in the image space using Approach II. From left to right, top to bottom, the figures are: 2212 blob-level scene events recognised from the testing set, (among which) 1116 "can taken" events, 33 "entering and leaving" events, 316 "shop keeper" events, 406 "browsing" events and 341 "paying" events

The learned mixture models were also utilised to recognise blob-level scene events online. The computational cost added by recognition was neglectable and the algorithm still ran at a speed of 6Hz. Although the parameters of mixture models were extracted from the training set, they were used for recognising events both in the training set and the testing set. For performance evaluation, the ground truth was labelled manually (see (a) and (b) of Fig. 6). The events recognition results at each frame are shown in (c), (d), (e) and (f) of Fig. 6. To achieve a degree of robustness in events 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 performance of our algorithm was measured using the detection rate and the false detection, which is the number of results without corresponding ground truth, for each class of event. Table 1 shows the results of autonomous events detection and classification using both approaches.



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

Table 1 Events detection and recognition results. "N" stands for "number of events"

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	85.7	100	0	0	10	100	100	0	0
Ent. & lev.	18	66.6	55.6	8	1	18	61.1	5.6	3	0
Shop keeper	12	75.0	66.7	1	0	12	33.3	50.0	1	1
Browsing	10	60.0	100	3	7	8	62.5	100	9	10
Paying	8	100	75.0	6	0	6	100	100	6	1

5 Conclusions

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 recognition 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 recognition of "browsing". A fusion of the two approaches could give more accurate recognition.

It was noticed that quite a lot of "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. Perhaps one should not expect the system to resolve this ambiguity unless higher order spatio-temporal correlations among different classes of events can 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.

To summarise, Pixel Change History (PCH) has been introduced as an effective representation for modelling autonomous visual events at the pixel-level. These pixel-level events are then used to automatically detect blob-level scene events. Our experiments show that such blob-level scene events can be given semantically meaningful interpretations. This is without the need for object trajectory matching. The work done so far only represents the first step toward a more comprehensive model for behaviour profiling and 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 based on autonomous pixel-level events.

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**: 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**: 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 E*1 4*NS*, 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. 该方法将场景事件与目标行为建模为一组使用 PCH(Pixel Change Histories)在像素级上检测的自治事件. 结合基于 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及手工标定的情况下自动识别场景级的事件的有效性.

关键词 自治事件检测, 行为造型, 期望最大化, 改进的最小描述长度, 运动历史图像, 象素变化历史, 场景事件, 轨迹匹配

中图分类号 TP391.41