Human Behaviour Analysis by using a Multi-Layered Finite State Markov Automaton

Thi Thi Zin, Pyke Tin, Takashi Toriu
Graduate School of Engineering
Osaka City University
Osaka, Japan
thithi@info.eng.osaka-cu.ac.jp

Hiromitsu Hama
Research Center for Industry Innovation
Osaka City University
Osaka, Japan

***Abstract*—Human behaviour analysis involving multiple person interactions has been motivated by the growing demand for recognising suspicious activity in security and surveillance applications. Especially, exchanging suspicious objects between persons is a common security concern in airports and other transit scenarios. In this context, we propose a new and efficient method for analysing multiple human behaviours and interactions based on a model of multi-layered finite state Markov Automaton. The model is built on a low-level image processing module for spatiotemporal detections and object tracking. The finite state Markov machine learns, in an unsupervised mode, usual patterns of behaviours in a scene over long periods. Then, in the recognition phase, usual behaviours are normal and deviant behaviour patterns are abnormal. Results, on real image sequences, demonstrate the robustness of the proposed method.**

Keywords-human behaviour; finite state Markov; suspicious activity; video surveillance; multiple person interactions

I. INTRODUCTION

Human behaviour analysis can be applied in a variety of application domains such as video surveillance, video retrieval, human-computer interaction systems and medical diagnoses. The results of such analysis can be used to identify people acting suspiciously and other unusual events directly from videos. Over the last decade, there has been growing interest within the computer vision and machine learning communities in the problem of analysing human behaviour in video. Such systems typically consist of a low or mid-level computer vision system to detect and segment a moving object-human or car based on Hidden Markov Models [1–4], Bayesian networks [5–6] and stochastic context free grammars [7].

However, there have been relatively few efforts to understand human behaviours that have substantial extent in time, particularly when they involve interactions between people. This level of interpretation is the goal of this paper, with the intention of establishing systems that can deal with the complexity of multi-person human behaviour in public places. Multiple person interactions have largely been motivated by the growing demand for recognising suspicious activity in security and surveillance applications. In [8], the behaviour detection process consists of foreground segmentation, blob detection and tracking. Semantic descriptions of suspicious human behaviour are defined through groups of low-level blob-based events. For example, fights are defined as many blobs' centric moving together, merging and splitting, and overall fast changes in the blobs' characteristics. Attacks are defined as one blob getting too close to another blob, with one blob perhaps being initially static, and one blob erratically moving apart [9–11].

On the other hand, interactions with objects that often include picking up an object, placing an object in the scene [12, 13] or passing an object to another person. Exchanging objects between persons is a common security concern in airports and other transit scenarios. This event was based on the localisation of spatiotemporal patterns of each human motion, and uses a shape and flow-matching algorithm. Analysing complex spatiotemporal changes in a dynamic scene in order to recognise logical sequences of usual activity patterns and detect unusual deviations has become an important problem in computer vision, especially in the context of visual surveillance [14–16].

In this paper, we present a new framework for human behaviour recognition based on a model of multi-layered Finite State Markov Machines (FSMM), built on top of a low-level image processing module for spatiotemporal detections and multiple object identification. We fix the modes of interaction in the lower physical layer of the network architecture depending on the context of the problem, and automatically find clusters of high probability sequences of transitions to learn usual behaviour patterns in an unsupervised manner. Low probability sequences, which are not recognised by the FSMM network, are diagnosed as unusual. Our approach to modelling person-to-person interactions is to use a FSMM learning techniques to teach the system to recognise normal single-person behaviours and common person-to-person interactions. A major emphasis of our work is on efficient Markov integration of both prior knowledge (by the use of synthetic prior models) and evidence from data (by situation-specific parameter tuning).

Our goal is to be able to successfully apply the system to any normal multi-person interaction situation without additional training. After the system has been trained at a few different sites, previously unobserved behaviours will be rare and unusual. To account for such novel behaviours, the system should be able to recognise new behaviours and to build models of them from as little as a single example. We have pursued a Markov automaton approach to modelling that includes both prior knowledge and evidence from data, believing that the

This work is partially supported by SCOPE: Strategic Information and Communications R&D Promotion Program (10103768).

Markov approach provides the best framework for coping with small datasets and novel behaviours [17].

The paper is structured as follows. Section II presents an overview of the system, which describes the computer vision techniques used for segmentation and tracking of the moving objects and the logical reasoning of Markov models used for behaviour modelling and recognition including human action and interaction events analysis. Section III contains experimental results with both synthetic agent data and real video data. Finally, Section IV summarises the main conclusions and sketches our future directions of research.

II. Multi-Layered Finite State Markov Automaton

Our architecture of the proposed multi-layered finite state Markov automaton for human behaviour analysis of complex actions is described in Fig. 1. It includes three layers: physical layer, logical layer and event layer. In the physical layer, two tasks will be performed: low-level image processing task and semantic interpretation task of physical events. The first task consists of the following modules:

- Segmentation using stochastic background subtraction,
- Identification of splitting and merging of image segments,
- Object tracking using finite state Markov process,
- Detection of the position of the object.

The semantic interpretation of the physical events is carried out in the FSMM's physical layer of the architecture. The interpretation at this level will be made in terms of the number of objects present in the scene, their split and merge history, their geometric positions, and changes in these positions with time.

Logical interpretations of the physical actions including high probability sequences of single object motions and multi-object interactions at the physical layer will be represented by the symbols and states in the logical layer. The types of interactions between objects that are to be considered are specified a priori depending on the context. The logical states are determined automatically as high probability sequences of physical transitions. Examples of such logical interpretations may be a person walking or running in a certain direction, a person walking toward a car, a person putting a bag down, two people moving toward each other etc. We then extract the high probability sequences in the logical layer as events, which are states in the event layer. Even more complex events that are characterised by high probability sequences of simpler events are detected at the event layer. Since the states at the higher layers are just high probability sequences at the lower layers detected automatically by clustering, recognition of usual behavioural patterns and detection of deviant ones can be carried out. In what follows we describe the details of each layer.

A. The Physical Layer

As shown in Fig. 2, the low-level image processing of the physical layer has first to carry out the tasks of super-pixel extraction and establishment of a super-pixel inference graph. The physical state production part includes the processes of super-pixel tracking and classification of state events. The details of each part are explained below.

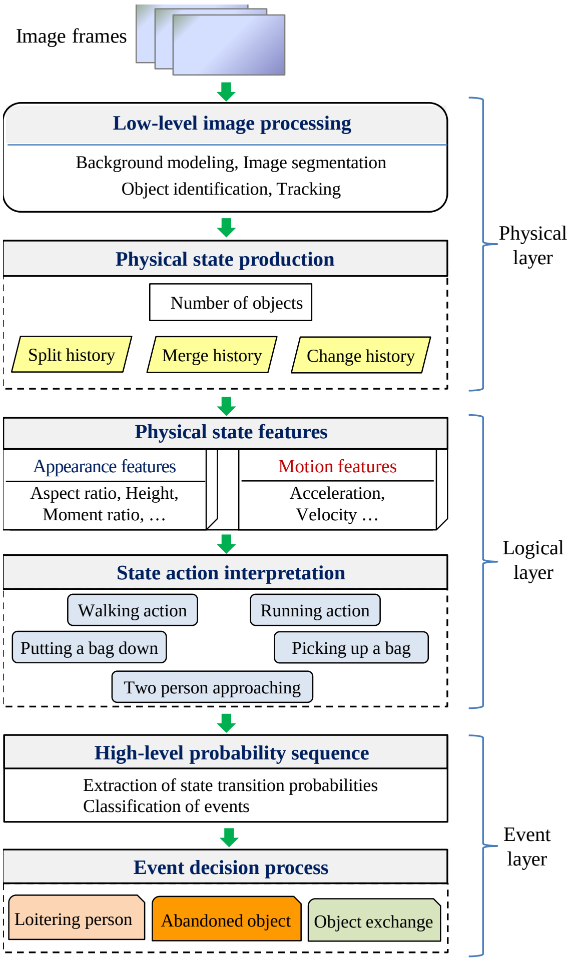


Figure 1. Overview of the proposed system.

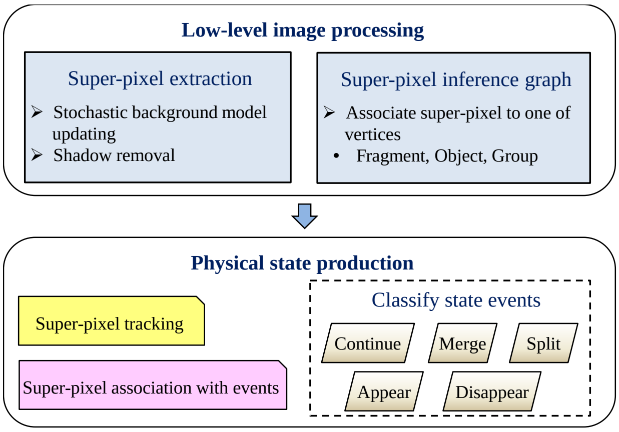

Low-level image processing
Super-pixel extraction
➢ Stochastic background model updating
➢ Shadow removal
Super-pixel inference graph
➢ Associate super-pixel to one of vertices
• Fragment, Object, Group
Physical state production
Super-pixel tracking
Super-pixel association with events
Classify state events
Continue
Merge
Split
Appear
Disappear

Figure 2. The process of physical layer.

1) Super-pixel Extraction: In this part, we segment an input image as foreground and background. Foreground is expected as objects of interest to be a target for tracking and action recognition. The ability to rapidly extract an object's correct appearance from an image is essential because object tracking and action modeling are based on it. To address this need, we develop a new stochastic background model so that we can effectively extract foreground in successive frames [18, 19]. As shown in Fig. 3(b), most of the results of foreground extraction include false positives of shadow and highlight induced by moving objects. To remove shadow and highlight from the result of foreground extraction, we assume that the intensity of a shadow pixel is scale-down of the intensity of the corresponding pixel in the background model, and we adopt the shadow and highlight detection as shown in Fig. 3(c). After removing shadow and highlight region, we get the final result of foreground extraction as shown in Fig. 3(d). Having obtained foreground pixels at frame *t*, through connected component analysis, they are clustered into a set of super-pixels $S^t = \{sp(i, t) \mid i$ is an integer and $0 \le i\}$ where a super-pixel $sp(i, t)$ is a i^{th} set of connected foreground pixels.

2) Object tracking: In the ideal case, a super-pixel represents an object, but in reality, multiple objects may appear as a single super-pixel (grouping), or an object may be broken into several super-pixels (fragmentation). To handle these problems, as shown in Fig. 4, we develop an online multiple objects tracking based on the Markov type framework. In particular, given S^t, we get a set of object tracks O^{t-k+1} at frame $t-k+1$ for $1 \le k < t$. Firstly, we detect super-pixel association events by associating S^t with S^{t-1}, and update the super-pixel inference graph according to super-pixel association events, and we label each vertex as one of Fragment, Object, and Group. Lastly, we localise objects using the super-pixel inference graph and its association event log.

3) Super Pixel Tracking: Given S^{t-1} and S^t extracted from two consecutive frames, we maintain super-pixel tracks by inferring super-pixel association events. As shown in Fig. 5 and Table I, super-pixel association events are classified into five events as follows: continue, merge, split, appear, and disappear. They can be inferred by a $|S^{t-1}|\times|S^t|$ correspondence matrix *P*. To make a correspondence matrix shown in Table II, we compare regions corresponding to S^{t-1} and S^t in which an element of *P* is set as follows:

$$p_{ij} = \begin{cases} 0 & \text{if } S_i^{t-1} \cap S_j^t = \phi, \\ 1 & \text{if } S_i^{t-1} \cap S_j^t = S_i^{t-1} \text{ or } S_i^{t-1} \cap S_j^t = S_i^t. \end{cases}$$

When *P* is determined, super-pixel association events are inferred as follows:

- Appear: S_j^t appears if $\sum_{i=1}^{|S_i^{t-1}|} p_{ij} = 0$,
- Continue: $S_j^{t-1} = S_j^t$ if $\sum_{i=1}^{|S_i^{t-1}|} p_{ij} = 0$ and $\sum_{j=1}^{|S_j^t|} p_{ij} = 1$,
- Merge: $\{S_i^{t-1}, p_{ij} = 1\}$ merge into S_j^t if $\sum_{i=1}^{|S_i^{t-1}|} p_{ij} \ge 0$,
- Split: S_i^{t-1} split into $\{S_j^t, p_{ij} = 1\}$ if $\sum_{j=1}^{|S_j^t|} p_{ij} \ge 0$.

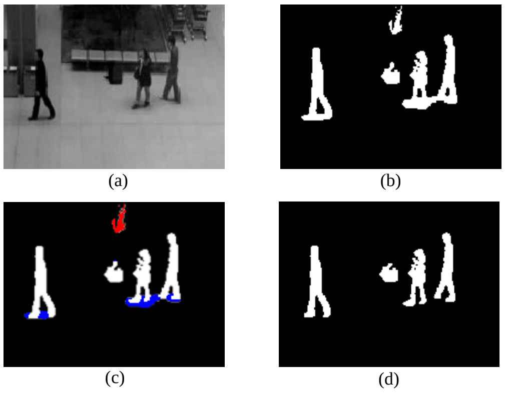


Figure 3. Effective foreground extraction: (a) an input image, (b) foreground extraction, (c) detected shadow (marked in blue) and highlight (marked in red), and (d) foreground extraction without shadow and highlight.

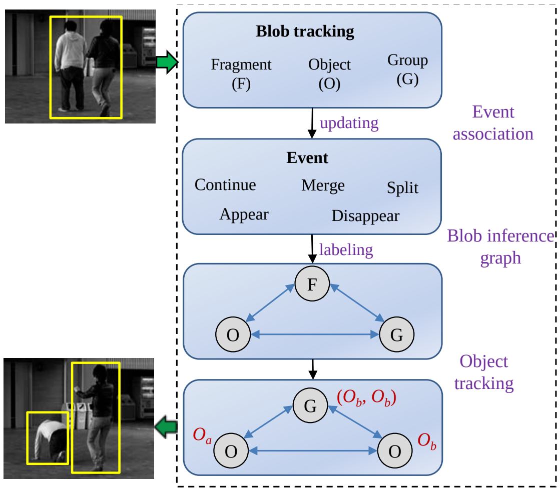


Figure 4. The multiple objects tracking framework.

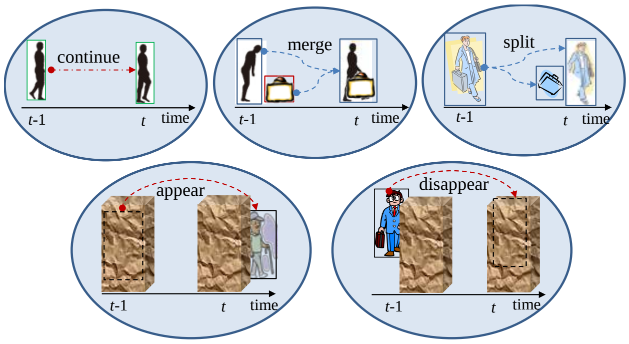


Figure 5. Super-pixel association state events.

TABLE I. SUPER-PIXEL ASSOCIATION STATE EVENTS

Events	**Associate states**		
	Time: *t*-1	⇒	***t***
Continue	S_1^{t-1}	⇒	S_1^t
Merge	$S_2^{t-1}+S_3^{t-1}$	⇒	S_2^t
Split	S_4^{t-1}	⇒	S_3^t, S_4^t
Appear		⇒	S_5^t
Disappear	S_5^{t-1}	⇒	

TABLE II. SUPER-PIXEL CORRESPODENCE MATRIX

		S_1^t	S_2^t	S_3^t	S_4^t	S_5^t
		(1)	(2)	(1)	(1)	(0)
S_1^{t-1}	(1)	1	0	0	0	0
S_2^{t-1}	(1)	0	1	0	0	0
S_3^{t-1}	(1)	0	1	0	0	0
S_4^{t-1}	(2)	0	0	1	1	0
S_5^{t-1}	(0)	0	0	0	0	0

After updating the inference graph according to the super-pixel association events, we label each vertex as one of Fragment (F), Object (O), and Group (G). We can then infer object tracks depending on the state change of the vertex label caused by super-pixel association events as shown in Fig. 6.

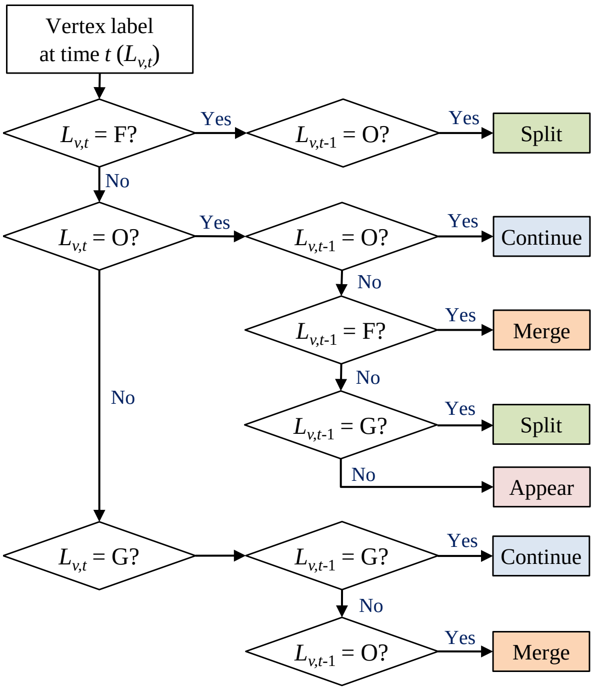


Figure 6. A flow chart for object association event inference ($L_{v,t}$ is the label of a vertex *v* at frame *t*).

B. The Logical Layer

In this layer, we represent the individual track pattern Γ (*i*, *t*) of the *i*th person at time *t* in terms of the features such as appearance features and motion features. The main interest in the track-level analysis includes the estimation of a moving person's speed, perimeter sentry for specific secured regions in a scene, the estimation of proximity between persons, etc.

$$\Gamma(i,j)=\left[P(i,j),d(i,j,t),|v_i|,\angle v_i\right]$$

$P(i,t)$: coordinates of the current track position

$d(i,j,t)$: relative distance between the i^{th} person and the most adjacent j^{th} person at frame *t*

$|v_i|$: track velocity magnitude

$\angle v_i$: track velocity orientation

The track-level analysis can be applied to moving persons, but it may not be effective to discriminate detailed human activity patterns performed by stationary people: e.g., turning, dropping an object, picking an object, shaking hands, dancing, pushing, kicking, etc. In this layer, we consider the high probability sequences of the physical layer to represent logical events. We extract all maximal sequences $S_n = i, j, \ldots, n$ such that $P(S_n) = P_{ij}P_{jk}\ldots P_{\bullet n} > Th_{\text{seq}}$ and $P_{lm} > Th_{\text{edge}}$ for every consecutive pair of states (l, m) in S_n and its corresponding transition probability P_{lm} , where Th_{seq} and Th_{edge} are the probability thresholds for a sequence and an edge respectively. We call such sequences High Probability Logical (HPL) sequences. Starting from any edge (i, j) , we develop an algorithm to find all sequences S_n with (i, j) as an edge by examining sub-sequences in both forward and backward directions. We ignore the backward branches of a sequence generated by all forward expansions in order to avoid reaching the same sequence from more than one path, as shown in Fig. 7. Every HPL sequence in the physical layer represents a state in the logical layer. The state transitions at the logical layer happen when one HPL sequence shifts to another.

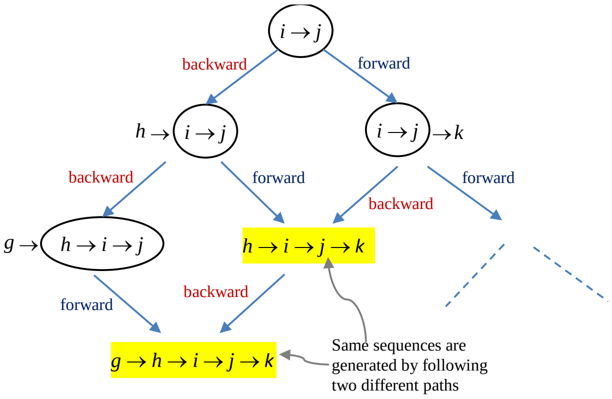


Figure 7. Example of multiple paths for the same sequence.

C. The Event Layer

For determining complex events that are high probability transitions in the logical layer, we employ an identical algorithm at the logical layer. In our current experiments, we have restricted our framework to three levels of hierarchy. During the training phase, the data is processed online to do the following tasks: (i) build generic FSMM's in the physical states as described in the logical layer and (ii) building of higher level states: with each transition in the physical state the HPL algorithm is run to see if any new logical states are found. Any such logical state is pushed up and stored for detection of unusual events.

We maintain a counter fail count with each HPL sequence, which tracks the number of times the FSMM corresponding to a HPL sequence fails to make a valid transition. A global counter, total count, maintains a count of the time from which any logical transition has occurred. If total count exceeds a certain threshold, then each fail count is checked. If all the fail counts are over a threshold, then the system is in an undefined logical state and an unusual event is flagged. If any transition in a HPL sequence is faster or slower than that predicted by the learning phase statistics, then an unusual time alarm is flagged.

Finally, an FSMM network described on the physical states provided by the lower level image processing layer can just be fed into the system for detecting unusual events in a supervised manner. In Fig. 8 and Fig. 9 we give one such example of the FSMM at the event layer corresponding to the sequence of logical events leading to the interpretation of two-person exchanging suspicious objects.

III. EXPERIMENTAL STUDY: VALIDATION AND EVALUATION

In this experimental study, we present an outdoor two-person interaction in a single-view. A set of atomic actions described in [20] is used as the basic building blocks of the activity hierarchy: i.e., enter, leave, walk leftward, walk rightward, bend down, stand up. These atomic actions are basic in that they do not depend on specific objects as targets, and are common to higher layers in the activity hierarchy. Object-handling activities form a category of more complex activities in single-person action and two-person interactions. Object exchange involves multiple persons as agents and forms the next layer in the activity hierarchy.

We consider various interaction types including direct exchange of objects between person 1 (P1) and person 2 (P2) and indirect exchange of objects where P1 leaves an object and P2 takes the object. We have built the FSMM based recognition system for activity patterns of a person. The states of the FSMM represent distance vectors between two persons, a person and an object, and events such as splitting and merging. The system is trained with sequences in which people walk-left, drop-object, pick-object, bend-up, bend-down, and walk-right. Twenty epochs were made by randomly choosing different sequences for the test data. The overall accuracy was 95%. In Fig. 9, we show an example image frames of the sequential actions where the two walking people interact with the objects. Usual events recognized are "walking persons carrying bags" and "dropping a bag". Unusual events (e.g. object hand-over) detected are "a person picking up another person's bag" and "walk with a different bag".

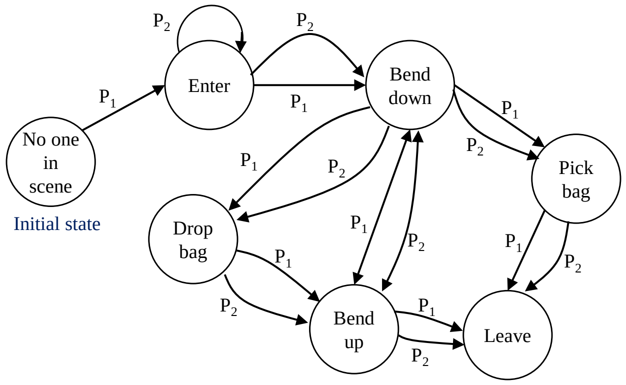
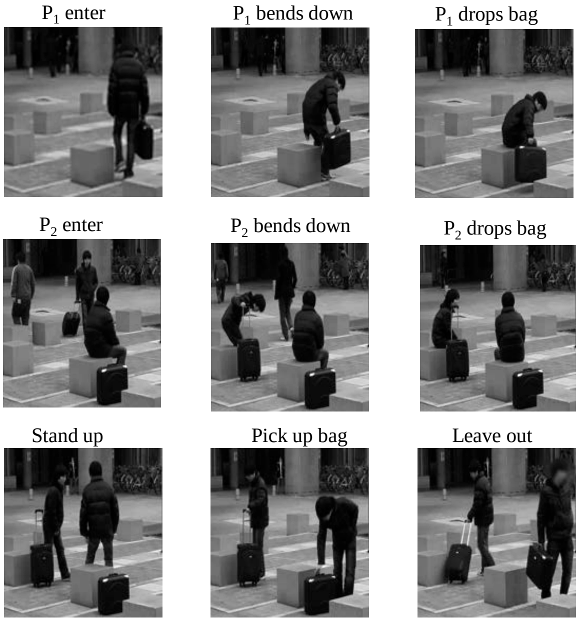


Figure 8. Example of FSMM in event layer.


P_1 enter

P_1 bends down

P_1 drops bag

P_2 enter

P_2 bends down

P_2 drops bag

Stand up

Pick up bag

Leave out

Figure 9. Some experimental results of object hand-over.

IV. CONCLUSIONS

We have presented a new framework for unsupervised learning of usual activity patterns and detection of unusual activities of human interactions based on a multi-layered FSMM. The FSMM framework is simple yet powerful and can reliably capture complex and concurrent inter-object interactions. Our preliminary results demonstrate the potential of the approach. In our present implementation, we have handled the cases of complex image processing and multiple objects tracking under occlusion. In the future, we plan to include such cases and test the framework rigorously under dense situations with many activities happening simultaneously.

ACKNOWLEDGMENT

We thank Scope project members and the students of Physical Electronics and Informatics, for their participation in producing tested videos.

REFERENCES

[1] B. Bose, X. Wang and E. Grimson, “Multi-class object tracking algorithm that handles fragmentation and grouping,” Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Minneapolis, USA, Jun. 2007.

[2] D. Arsic, B. Schuller, and G. Rigoll, “Suspicious behavior detection in public transport by fusion of low-level video descriptors,” Proc of 8th Intl. Conf. on Multimedia and Expo (ICME 2007), Beijing, China, pp. 2018–2021, Jun. 2007.

[3] Z. Zhang and M. Piccardi, “A review of tracking methods under occlusions,” Proc. of the IAPR Conf. on Machine Vision Applications (IAPR MVA 2007), Tokyo, Japan, pp. 146–149, May. 2007.

[4] N. Oliver, E. Horvitz, and A. Garg, “Layered representations for human activity recognitio*n*,” In 4th IEEE Intl. Conf. on Multimodal Interfaces, pp. 3–8, 2002.

[5] S. Hongeng and R. Nevatia, “Multi-agent event recognition,” Proc. of Intl. Conf. on Computer Vision, pp. 84–93, 2001.

[6] H. Zhong , J. Shi and M. Visontai, “Detecting unusual activity in video,” Proc. of Intl. Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 819–826, 2004.

[7] Y. Ivanov and A. Bobick, “Recognition of visual activities and interactions by stochastic parsing,” IEEE Trans. On Pattern Anal. and Machine Intell., 22(8), pp. 852–872, Aug. 2000.

[8] B. Morris and M. Trivedi, “Learning trajectory patterns by clustering: Experimental studies and comparative evaluation,” IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 09), Miami, USA, pp. 312–319, Jun. 2009.

[9] L.M. Fuentes and S.A. Velastin, “Tracking-based event detection for CCTV systems,” Pattern Analysis and Applications, vol. 7, no. 4, pp. 356–364, 2005.

[10] H. Yasin and S.A. Khan, “Moment invariants based human mistrustful and suspicious motion detection, recognition and classification,” Proc. of 10th Intl. Conf. on Computer Modeling and Simulation, Cambridge, UK, pp. 734–739, 2008.

[11] S. Park and J. K. Aggarwal, “Recognition of two-person interactions using a hierarchical Bayesian network,” First ACM SIGMM International Workshop on Video Surveillance, ACM Press, pp. 65–76, 2003.

[12] Thi Thi Zin, H. Hama, Pyke Tin and T. Toriu, “Evidence fusion method for abandoned object detection,” ICIC Express Letters (Part B: Applications), vol. 2, no. 3, pp.535–540, Jun. 2011.

[13] Thi Thi Zin, Pyke Tin, H. Hama and T. Toriu, “Unattended object intelligent analyzer for consumer video surveillance,” IEEE Trans. on Consumer Electronics, Vol. 57, No. 2, pp. 549–557, May. 2011.

[14] J. Kim and K. Grauman, “Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates,” IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Miami, USA, pp. 2921–2928, 2009.

[15] L. Kratz and K. Nishino, “Anomaly detection in extremely crowded scenes using spatiotemporal motion pattern models,” IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Miami, USA, pp. 1446–1453, 2009.

[16] Thi Thi Zin, Pyke Tin, T. Toriu and H. Hama, “A Markov random walk model for loitering people detection,” In the 6th Intl. Conf. on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP-2010), Darmstadt, Germany, pp. 680–683, Oct. 15-17, 2010.

[17] Pyke Tin, Thi Thi Zin, T. Toriu and H.Hama, “A general framework for knowledge based human behavior understanding,” The Sixth International Conference on Innovative Computing, Information and Control (ICICIC2011), Kitakyushu, Japan, Dec. 22-24, 2011, in press.

[18] Thi Thi Zin, Pyke Tin, T. Toriu and H. Hama, “Background modeling using special type of Markov chain,” IEICE Electronic Express, Vol. 8, No. 13, pp.1082–1088, Jul. 2011.

[19] Thi Thi Zin, Pyke Tin, T. Toriu and H. Hama, “An innovative background model based on multiple queuing framework,” The Sixth International Conference on Innovative Computing, Information and Control (ICICIC2011), Kitakyushu, Japan, Dec. 22-24, 2011, in press.

[20] M. Sugimoto, Thi Thi Zin, T. Toriu and S. Nakajima, “Robust rule-based method for human activity recognition”, International Journal of Computer Science and Network Security, Vol. 11, No. 4, pp. 37–43, Apr. 2011.