International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:11, No:9, 2017

Human Behavior Modeling in Video Surveillance of Conference Halls

Nour Charara, Hussein Charara, Omar Abou Khaled, Hani Abdallah, Elena Mugellini

***Abstract*—**In this paper, we present a human behavior modeling approach in videos scenes. This approach is used to model the normal behaviors in the conference halls. We exploited the Probabilistic Latent Semantic Analysis technique (PLSA), using the 'Bag-of-Terms' paradigm, as a tool for exploring video data to learn the model by grouping similar activities. Our term vocabulary consists of 3D spatio-temporal patch groups assigned by the direction of motion. Our video representation ensures the spatial information, the object trajectory, and the motion. The main importance of this approach is that it can be adapted to detect abnormal behaviors in order to ensure and enhance human security.

***Keywords*—**Activity modeling, clustering, PLSA, video representation.

I. Introduction

DURING a conference, the sequence of events is almost unified and standardized; it has a defined set of activities. Based on these observations, the normal model is constructed by unsupervised learning of all the activities expected to be normal with respect to the different depth zones, following a single-class learning mode. The unsupervised learning process of the model aims to group videos with similar activities or patterns into the same group topic. All of these subjects, which are learned as normal, will be assigned to an encompassing class, the normal class. Note that the definition of unsupervised learning prevents labeling or supervision on the data. Let us now discuss the method of unsupervised learning of the different subjects. Conventional clustering methods are most commonly used to learn groups of unlabeled data. These methods aggregate the data using a distance measure defined on the feature space. However, the choice of distance measurement plays a critical role in determining the shape, size and quality of clusters. On the other hand, the ability of Probabilistic Topic Models (PTM) to group co-occurring words encourages us to compare them with classical clustering methods. PTMs avoid the complicated selection of distance measurements between visual characteristics, as in conventional methods, which resort to co-occurrences of characteristics. Furthermore, in PTM models, grouping keys are based on a vocabulary of words and can be extended to the temporal criteria of documents. Thus, we adopt a method based on the models of probabilistic subjects in our system and we validate a simple hypothesis which states that the activities leave a trace of the co-occurring observations. Then, we exploit the PLSA model belonging to the PTM family and that was originally developed for) the behaviors in the videos, PLSA adopts the paradigm 'Bag-of-Words' which consists of the extraction of characteristics and the construction of a dictionary adapted to each context. Similarly, the representation of the behavior and the extracted characteristics are adopted. In particular, several issues must be taken into account when choosing methods of representation. First, it is necessary to know the specific behaviors that occur. Second, it is necessary to select the characteristics that are capable to isolate the salient properties. Third, according to the tool of the activity analysis chosen, we should study how to take into account the characteristics that are quantified in discrete compartments (the 'terms'). Finally, how the temporal dynamics of behavior can be exploited. Indeed, the PLSA model using the 'bag-of-terms' paradigm ignores the temporal order of terms in a video clip; hence, the need for a representation that integrates the temporal information into the model.

N. Charara is a Professor at American University for Culture and Education, Computer Science Department, Beirut, Lebanon (e-mail: nourchrara@auce.edu.lb n, charara@hotmail.com).

H. Charara is a Professor at Lebanese University, Faculty of Science. Nabatieh, Lebanon (e-mail: hussein.charara@ul.edu.lb, h_charara@hotmail.com).

Hani Abdallah is the dean of the Arts and Science faculty at American University for Culture and Education, Computer Science department, Beirut, Lebanon (e-mail: deanofscience@auce.edu.lb).

Elena Mugellini and Omar Abou Khaled are with the ICT Department, University of Applied Sciences of Western Switzerland, Perolles 80, CH-1705 Fribourg, Switzerland (e-mail: name.surname@hefr.ch).

II. Related Works

A typical video surveillance system for behavior recognition normally explores various video processing levels. The objective is to analyze the monitored scene and extract useful information in order to finally arrive at a semantic interpretation for a specified application. The behavioral representation aims to isolate the prominent properties of the video data, in particular those that allow useful discrimination between interesting events. The various works performed for the representation of behavior or abstraction [2] can be grouped according to the scale dealing with the problem of description of the data on a microscopic or macroscopic scale.

A. Microscopic Approaches

First, the most basic form is the estimation of the foreground pixels [3], [5]. The second form is well known in the understanding of behaviors, which is the optical flow [8], [4]. This generally involves extracting movement (direction and speed) from the individual pixels between the consecutive frames. The characterization according to the appearance of image is the third form of the microscopic representation; it has presented encouraging results in recent studies [1]-[6].

B. Macroscopic Approaches

Macroscopic representation is an abstraction approach based on the assumption that the description of individual objects participating in the video sequence is a good intermediate representation for event reasoning. Thus, a set of low-level features for each segmented object is constructed, including trajectory, velocity, shape, and blob descriptors. Such methods are used for example for the detection of a fall of an object from the shape and the history of the movement [7]-[9], the 3D tracing of the shape of a human head [11] Or the processing of the 2D - 3D alignment problem in the detection of the type of object with variation of point of view [10].

C. Mixed Approaches

Instead of using only a microscopic or macroscopic representation, some work has benefited from exploiting the advantage of each of these two approaches, supporting a mixed approach.

Lin et al. [12] describe a video surveillance system based on pixel-level characteristics. In particular, color, distance, and a count function, where evolution techniques are used to measure the similarity of observations. At the object level, the system follows each person and classifies their behavior by analyzing their trajectory models. This is done with a hybrid genetic algorithm that uses a Gaussian synapse. In the work of Neverova et al. [13], the overall appearance of each gesture instance is captured by the skeleton descriptor, while the video streams convey additional information about the hand shapes and their dynamics. This representation on two scales is essential for the discrimination between the classes of gestures performed in poses of similar bodies. Some works use criteria at different levels in order to expand the range of detected events. For example, Varadarajan and Odobez [15] considered the orientation of the optical flow in each pixel (north, south, east, west) and the position of this pixel in the scene.

For behavior modeling techniques, graphic models provide a flexible framework for modeling large collections of variables with complex interactions. Probabilistic graphic models deal with a variety of models, ranging from Bayesian networks, dynamical Bayesian networks, their extensions and random Markov fields [14], [16], [17]. Other than the Bayesian networks, different models have been proposed for the modeling of behaviors. In this context, Varadarajan et al. [18] propose a new graphic model called Mixed Event Relationship (MER), which integrates learning of local rules and global states simultaneously from a binary event matrix. PTMs have recently been applied to behavioral analysis. The PTMs are basically bag-of-words patterns that perform clustering by the occurrence, i.e. the subject is a group of co-occurring words. Mehran et al. [19] used the Dirichlet Latent Allocation (LDA) model to model human interactions within the crowd. Li et al. Use the LDA model in [20], respectively. Kooij et al. [21] proposed a model of subject, where they refer each behavior to a subject. Each behavior is an SLDS (Switching Linear Dynamics Systems), which defines by a LDS (Linear Dynamic System) the spatial location and the motion dynamics for each common action.

III. Video Representation

We adopt a microscopic representation at the pixel level since it can preserve most of the information for the recognition of the activity. The video is thus represented by separating the elements of interest (people) from the background and processing them in the form of the spatio-temporal patches. The characteristic of the movement accompanied by these space-time cubes is used to study the dynamics of the persons in the scene. The extraction and representation procedures are used for the construction of our vocabulary (Section V). These procedures involve segmenting the foreground elements (people in the normal case) and extracting a motion descriptor. They are chosen and designed to involve the trajectory information and the shape during the analysis generated by the PLSA.

Foreground Pixel Detection: We used an improved background modeling technique that is proposed by Klare and Sarkar in [22]. They adapted a conventional approach that models the pixel by a Gaussian, but instead of using Gaussian based on RGB intensity only, they used 13 Gaussians to represent a pixel. This technique is therefore able to manage the lighting of the scene.

Motion extraction: We try to describe the movement of the separated foreground objects. To distinguish between static and moving regions on the one hand, and to identify the direction of motion of mobile regions on the other hand, we use an optical flow estimation method. The estimation method used in our system is the differential method of Lucas and Kanade [23]; it is a very old but widely used method.

IV. Model Learning Using PLSA

The basic technique used to learn the behavior pattern is to treat the videos as textual documents. The extracted features that represent the behaviors of the video are treated as terms by producing a discrete description for each video document. Therefore, we get a set of terms, and a set of documents containing these terms. Then, PLSA is applied to these documents to obtain the probability models of the subjects; these are used to classify any new video. The semantic analysis generated by PLSA studies the spatio-temporal relations between the terms throughout the documents, in order to implicitly estimate the shape, the trajectory and the direction of the elements of interest.

Let $D = \{d_1, \dots, d_N\}$ be the set of video sequences with vocabulary terms $V = \{t_1, \dots, t_M\}$, in which the sequential order of the appearance of the terms is ignored. And let Z $Z = \{z_1, \dots, z_K\}$ be the set of unobserved subjects called latent subjects. In our application, a latent subject corresponds to a model of activity.

The data of D and V can be summarized by an $N \times M$ matrix of co-occurrence $\mathrm{N} = [n(d_i, t_j)]_{ij}$ with $i \in \{1, 2, \dots, N\}$ and $j \in \{1, 2, \dots, M\}$, and $n(d_i, t_j)$ is the number of occurrences of the term t_j in document d_j. This matrix is called a term-document matrix.

Suppose further that with each observation $\langle d_i, t_j \rangle$, a latent variable $z_k \in Z$ is associated.

In the process described above, the conditional probability $P(z_k|\, d_i)$of the subject z_k given the document d_i describes to what extent the document d_i explains the latent subject z_k. In other words, the importance given to each subject z is given by the probability distribution $p(z|\, d)$. Similarly, the conditional probability $P(t_j|z_k)$ of the term t_j knowing that the latent subject z_k tells us the contribution of the term t_j in the explanation of the subject z_k.

Using the conditional independence hypothesis in the model, the generation of a term t in a document d can be translated by the following joint likelihood model:

$$P(d_i, t_j) = P(d_i)\, P(t_j|d_i) \tag{1}$$

with

$$P(t_j|d_i) = \sum_{k=1}^{K} P(t_j, z_k|\, d_i) = \sum_{k=1}^{K} P(t_j|\, z_k)\; P(z_k|d_i) \tag{2}$$

Hence,

$$P(d_i, t_j) = P(d_i) \sum_{k=1}^{K} P(z_k|\, d_i)\; P(t_j|z_k) \tag{3}$$

Training Phase: For adjusting the PISA model, we must learn unobservable probabilities are the parameters of the model $\Xi = \{P(t_j|\, z_k), P(z_k|d_i): t_j \in T, d_i \in D, z_k \in Z\}$. These probabilities are estimated iteratively using the maximal likelihood principle, and the possible hidden aspect is deduced therefrom. More precisely, given a set of training documents, the logarithmic likelihood and likelihood of model parameters Ξ can be expressed as L and $\mathcal{L}$ given by formulas (4) and (5), respectively.

$$L = \prod_{i=1}^{N} \prod_{j=1}^{M} P(d_i, t_j)^{n(d_i, t_j)} \tag{4}$$

$$\mathcal{L} = log\, L = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, t_j)\, log\; P(d_i, t_j) \tag{5}$$

By replacing $P(d_i, t_j)$given by (6) in $\mathcal{L}$ and applying the basic properties of the logarithm function, $\mathcal{L}$ becomes:

$$\mathcal{L} = \log L = \sum_{i=1}^{N} n(d_i) \left[log\; P(d_i) + \sum_{j=1}^{M} \frac{n(d_i, t_j)}{n(d_i)} log\left[\sum_{k=1}^{K} P(z_k|d_i)\, P(t_j|z_k)\right]\right] \tag{6}$$

The optimization is performed according to the Expectation-Maximization (EM) algorithm [H01]. After initializing the parameters of the model Ξ using the random values, the EM procedure repeats two steps until convergence. In the Expectation E step, a posteriori probabilities for the latent variables are estimated, given the observations and using the current parameter estimates.

$$P(z_k|d_i,\; t_j\;) = \frac{P(t_j, z_k|d_i)}{P(t_j|d_i\;)} = \frac{P(t_j|z_k, d_i) P(\; z_k|d_i)}{P(t_j|d_i)} = \frac{P(t_j|z_k) P(z_k|d_i)}{\sum_{l=l}^{K} P(t_j|z_l) P(z_l|d_i)} \tag{7}$$

In the Maximization M step, the parameters are updated by maximizing logarithmic likelihood data ($\mathcal{L}$) and using the posterior probabilities of step E:

$$P(\; t_j|z_k,\;) = \frac{\sum_{i=1}^{N} n(d_i, t_j) P(z_k|d_i, t_j)}{\sum_{m=1}^{M} \sum_{i=1}^{N} n(d_i, t_m) P(z_k|d_i, t_m)} \tag{8}$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i, t_j) P(z_k|d_i, t_j)}{n(d_i)} \tag{9}$$

$P(d_i)$, $n(d_i)$ and $n(d_i, t_j)$can be directly estimated from the data. Step E (7) and step M ((8) and (9)) must be applied until the convergence condition is satisfied (6).

Classification Phase: After the training of the model and during the tests, we execute PLSA for all the video tests, applying the EM algorithm again without updating $P(t_j|z_k)$. The learned model $P(t_j|z_k)$ is used to learn $P(z_k|\, d_{test})$ and maximize the logarithmic likelihood of a document:

$$LL(d_{test}) = \sum_{j=1}^{M} n(d_{test}, t_j)\; log\left(\sum_{k=1}^{K} P(z_k|\, d_{test}) P(t_j|z_k)\right) \tag{10}$$

The complexity of PLSA is $O(K \times ŋ)$, where ŋ is the number of pairs (t, d)occurring at least once in the collection (or in other words the number of non-zero entries in the matrix Term-document).

V. Vocabulary and Documents Construction

The definition of a semantic space is essential to learn the different models of activity (the subjects) in an unsupervised way. Within the framework of PLSA, this space consists of all the documents and the entire vocabulary.

Documents: A document in our system consists of a short video clip and plays the role of a bag-of-terms. To generate the documents, the input video is divided by the time into short clips of n_t frames. These short clips overlap in n_t /10 frames to ensure continuity of information. Thus, these documents (short clips) will undergo a grouping according to the similarity and the frequency of co-occurrence of the terms included therein. Note that the value of n_t plays an important role in the operation of the system and controls accuracy and continuity and coverage.

Vocabulary: The most important aspect that controls the functioning of our system is the choice of the terms that describe the videos. Indeed, we must grab the forms of information that are able to appear and regroup throughout the documents, to give us the expected topics.

The activities are normally done by people. These people belong to the foreground, these foreground elements will then be detected and segmented as already presented in section III. Once people are segmented and separated, the trajectory, shape and direction information characterizing the activities of these people are extracted to construct the terms of PLSA.

The terms are constituted successively by two versions. The first version consists of 3D spatio-temporal patch groups. The second version is obtained by assigning the direction of motion to each of these spatio-temporal patches

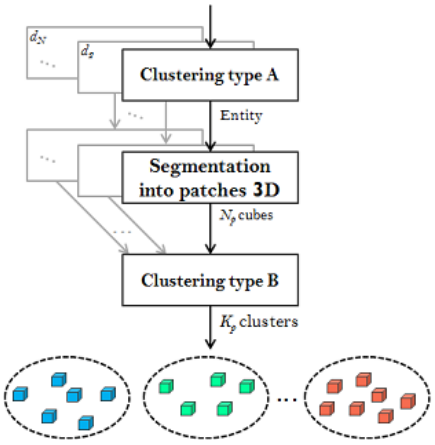

Fig. 1 The generation of the spatio-temporal patch groups of the first version

A. The First Version: 3D Spatio-Temporal Patch Groups

We first discuss the generation of the spatio-temporal patch groups of the first version (Fig. 1). The foreground pixels are the building blocks of these 3D patches. First, we try to group these pixels into clusters at the level of each document. Each of the clusters of considerable size represents an entity, which is in this application only a person with a continuous movement or appearance. Then, the pixels of these entities are segmented into 3D space-time cubes or patches. These spatio-temporal patches undergo a second grouping along all the documents to obtain the patch groups of the first version.

Clustering type A: Consider, for each document (short clip), a three-dimensional space that contains the pixels of foreground. This space-time volume is formed by the n_t consecutive foreground frames and distributed over the time axis. The data can be represented in the form of a matrix $N_F \times 3$ containing all the coordinates of the foreground pixels obtained by segmentation of which 3 is the number of coordinates (x, y, t) and $N_F$is the Number of the foreground pixels detected $f_i(x_i, y_i, t_i)\ \forall i \in \{1, 2, \dots, N_F\}$ in a document ie. On the n_t frames. Each row of the matrix thus corresponds to the coordinates of a pixel f_i. The neighborhood criterion between these pixels is explored in order to group them into independent clusters. This grouping step, called 'grouping of type A', plays two roles at this level. First, by looking at the volume of each cluster (the number of pixels belonging to the cluster), clusters of very small sizes are considered noise and are excluded in the next steps. In this way, we filter the erroneous results that often occur by all separation methods and are due to thresholding sensitivity. Second, this grouping allows separating and defining the different entities.

Segmentation into 3D patches: Let Υ be the space made up of the N_R pixels of foreground $p_k(x_k, y_k, t_k)$, $\forall k \in \{1, 2, \dots, N_R\}$ in a document after elimination of small clusters. We define a cubic patch $P_{(\Delta x, \Delta y, \Delta t)_j}$by a region of fixed size $\Delta x \times \Delta y \times \Delta t$, fully characterized by its origin point (x_j, y_j, t_j) By the pixels belonging to $\{(\alpha, \beta, \gamma) : \alpha \in (x_j, x_j + \Delta x), \beta \in (y_j, y_j + \Delta y), \gamma \in (t_j, t_j + \Delta t)\}$ and of which more than N_c fractions of its pixels, including its point of origin, belong to the space Υ. Cubic patches do not overlap and are generated to cover all pixels of Y, i.e. each pixel $p_k(x_k, y_k, t_k) \in \Upsilon, k \in \{1, 2, \dots, N_R\}$) must belong to a single element of the set $S_{P_{(\Delta x, \Delta y, \Delta t)}} = \left\{P_{(\Delta x, \Delta y, \Delta t)_j}, \forall j \in \{1, 2, \dots, N_P\}\right\}$, N_P is the total number of cubic patches generated in a document.

Grouping type B: The first version of the vocabulary is constructed by grouping all the cubic patches generated along all the documents. This grouping, called grouping of type B, is done at the point of origin (x_j, y_j, t_j) to arrive at a dictionary of $K_p$clusters common to all documents.

B. The Second Version: Integration of the Direction of Movement

The final version of the vocabulary is obtained by a scalar product of the different cubic patch clusters (characterized by their original points) with the motion space (Fig. 2). This space is constructed by quantifying the optical flow in three or four bins. Indeed, the optical flux is calculated for each pixel of foreground along all the documents. The use of a low threshold at the amplitude of the flux vector separates static pixels from those in motion. They are still categorized by quantifying their direction of movement, either in two classes (right-hand movement and left-hand movement) or in three classes (right-hand movement, left-hand movement and vertical movement). Indeed, the intention behind quantification in only two categories is to ignore the vertical secondary movements (e.g. the movement of the presenter's arms), especially since the nature of the normal activities taking place in our scenes is intended to it is horizontal, in both directions. On the other hand, this horizontal quantification may ignore the vertical abnormal movements by forcing them to assume a horizontal nature and consequently consider them normal. This depends on the influx of the location characteristic (the cubic patches) in the operation of the system. Our tests are therefore carried out for these two choices.

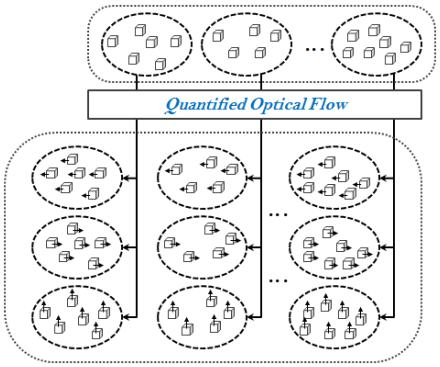

Fig. 2 Integration of the motion direction (left, right and static) to get the final version of the vocabulary

The optical flux values for the foreground pixels of each cubic patch are averaged. Thus, in total, we will have $3 \times$

$K_p$terms (or $4 \times K_p$ terms) in our dictionary.

VI. Learning Process

After obtaining the dictionary of terms, the cubic patches in each document are mapped to one of the $3 \times K_p$ common terms (or $3 \times K_p$ terms) of the dictionary, to generate the *term-document* co-occurrence matrix (Fig. 3). The matrix thus obtained is treated by PLSA to find spatiotemporal correlations between these patches in order to group them into K subjects, and consequently, to define the K types of activity.

From the *term-document* matrix, we can discover the different latent classes z, which capture the direction and direction of the cubic patches as well as their spatial arrangement. Thus, each of the documents (video clips) can be grouped into one of the hidden subjects using the Expectation Maximization algorithm, where steps E and M alternate by applying (7) (in Step E), (8) and (9) (in step M). After convergence, the algorithm calculates the probability of belonging each document and each term to each of the K hidden subjects.

Documents and similar words are those that have a high likelihood of belonging to the same subject. In fact, a subject z_k ($z_k \in Z = \{z_1, \dots, z_K\}$is assigned to a term if the associated activity model A_k (11) consists of the set of terms that are frequently occurring with this subject (12).

$$A_k = \{ t_j / j \in \{1, \dots, 3K_p\} \; and \; subject(t_j) = k\} \quad (11)$$

$$subject\,(t_j) = argmax_{z \in Z} \{P(t_j | z_k)\} \quad (12)$$

VII. Conclusion

We have presented a behavior modeling approach of conference hall video scenes.

PLSA is used to analyze activity and learn normal patterns in a lecture. The activity models are constructed by unsupervised learning by estimating the parameters of PLSA by a generative process. The bag-of-terms principle adopted by PLSA makes modeling performance largely dependent on the choice of terms that are used to capture the information contained in the documents.

Thus, an appropriate choice of words is essential to obtain good models and well separated clusters. Our representation on a microscopic scale is based on the separation of the elements of interest (the people) from the background and the study of their dynamics by the extraction of the motion, using a technique of generation of the spatio-Time. The characteristic of the movement accompanied with these space-time cubes is used to study the dynamics of the people in the scene. Thus, this representation preserves most of the information which characterizes activity and which distinguishes between its different kinds and at the same time implicitly assures the characteristics of the trajectory, speed, and form. As a future work, the patterns obtained by PLSA can be used to detect abnormal events.

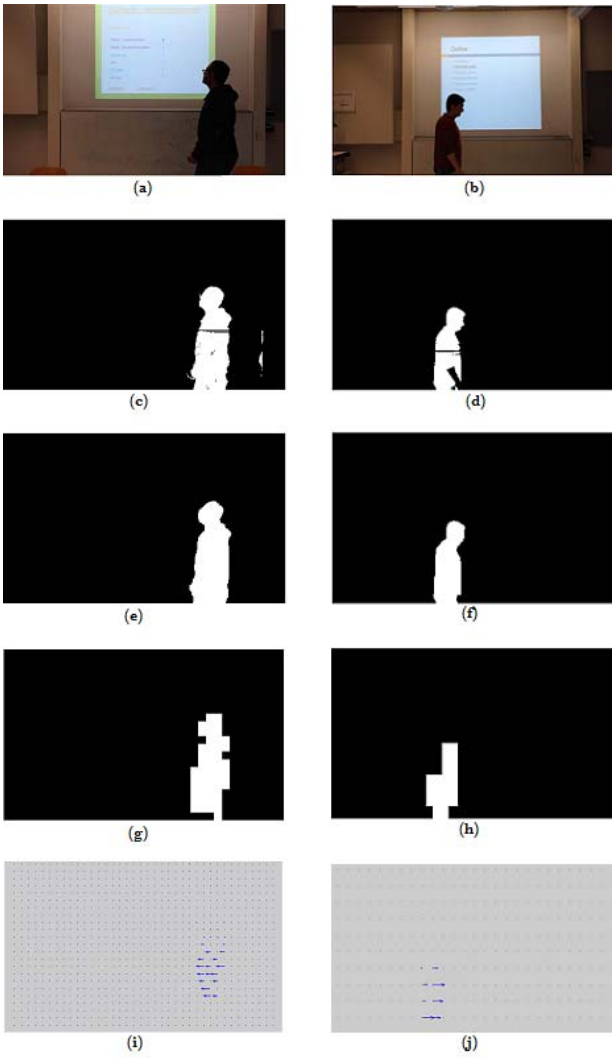

Fig. 3 Visualization of the output of each step of the construction of the vocabulary, taking an example in each column: (a-b) Original frames; (c-d) Maps of foreground pixels; (e-f) Maps of filtered and filled foreground pixels; (g-h) 2D Cube Pattern map; (i-j) Motion

References

[1] A. Adam, E. Rivlin, I. Shimshoni et D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(3):555–560, 2008.

[2] A. Torralba, K. P. Murphy et W. Freeman. Contextual models for object detection using boosted random fields. In Advances in neural information processing systems (NIPS). MIT Press, Cambridge, MA, pp 1401–1408, 2004.

[3] V. Reddy, C. Sanderson, A. Sanin, B et C. Lovell. MRF-based Background Initialisation for Improved Foreground Detection in Cluttered Surveillance Videos. CoRR abs/1406.5095, 2014.

[4] E. L. Andrade, S. Blunsden et R. B. Fisher. Modelling crowd scenes for event detection. Int. Conf. Pattern Recognition, Washington, DC, pp. 175–178, 2006.

[5] P.-M. Jodoin, J. Konrad et V. Saligrama. Modeling background activity for behavior subtraction. In International Conference on Distributed Smart Cameras, pages 1–10, 2008.

[6] C. Li, Z. Han, Q. Ye et J. Jiao. Abnormal behavior detection via sparse reconstruction analysis of trajectory. In the 6th International Conference on Image and Graphics (ICIG '11), pp. 807–810, 2011.

[7] C. Rougier, J. Meunier, A. St-Arnaud et J. Rousseau. Fall detection from human shape and motion history using video surveillance. In Proceedings Advanced Information Networking and Applications Workshops, 2007.
[8] S. Ali et M. Shah. A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis, in Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1–6, 2007
[9] A. Nasution et S. Emmanuel. Intelligent video surveillance for monitoring elderly in home environments. In IEEE Workshop on Multimedia Signal Processing, 2007.
[10] M. Aubry, D. Maturana, A. Efros, B. Russell et J. Sivic. Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models. In CVPR 2014.
[11] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau. Monocular 3d head tracking to detect falls of elderly people. In Proceedings IEEE Conference of the Engineering in Medicine and Biology Society, 2006.
[12] L. Lin, Y. Seo, M. Gen et R. Cheng. Unusual human behavior recognition using evolutionary technique. Computers and Industrial Engineering 56, 1137-1153, 2009.
[13] N. Neverova, C. Wolf, G. W. Taylor, F. Nebout. ModDrop: adaptive multi-modal gesture recognition. Minor revision at IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014.
[14] M. Sivarathinabala et S. Abirami. An Intelligent Video Surveillance Framework for Remote Monitoring. International Journal of Engineering Science and Innovative Technology (IJESIT). Volume 2, Issue 2, 2013..
[15] J. Varadarajan et J.M. Odobez. Topic models for scene analysis and abnormality detection. In Proceedings of the International Conference on Computer Vision - Workshop on Visual Surveillance, Kyoto, 2009.
[16] Y. Wang, D. Wang et F. Chen. Abnormal Behavior Detection Using Trajectory Analysis in Camera Sensor Networks, International Journal of Distributed Sensor Networks, Article ID 839045, 9 pages, 2014.
[17] R. Ge, Z. Shan, H. Kou. An Intelligent Surveillance system Based on Motion detection, Proceedings of IEEE, IC-BNMT 2011.
[18] J. Varadarajan, R. Emonet, et J. Odobez. Bridging the Past, Present and Future; Modeling Scene Activities from Event Relationships and Global Rules. In IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012.
[19] R. Mehran, A. Oyama et M. Shah. Abnormal crowd behaviour detection using social force model. In IEEE Conference on Computer Vision and Pattern Recognition, pages 935–942, 2009.
[20] J. Li, S. Gong, et T. Xiang. Discovering multi-camera behaviour correlations for on-the fly global activity prediction and anomaly detection. In IEEE International Workshop on Visual Surveillance, Kyoto, Japan, 2009.
[21] J. F. P. Kooij, G. Englebienne, D. M. Gavrila. A Non-parametric Hierarchical Model to Discover Behavior Dynamics from Tracks, Computer Vision – ECCV 2012 Lecture Notes in Computer Science, volume 7577, pp 270-283, 2012.
[22] B. Klare et S. Sarkar. Background subtraction in varying illuminations using an ensemble based on an enlarged feature set. Computer Vision and Pattern Recognition Workshop, 0:66–73, 2009.
[23] B.D. Lucas et T. Kanade. An iterative image registration technique with an application to stereo vision. In Proceedings of the International Joint Conference on Artificial Intelligence, pages 674–679, 1981.