Localized Anomaly Detection via Hierarchical Integrated Activity Discovery

Thiyagarajan Chockalingam[1]
thambu2003@live.in

Rémi Emonet[2]
http://home.heeere.com

Jean-Marc Odobez[2,3]
odobez@idiap.ch

[1]: Colorado State University – Fort Collins, CO 80523, United States
[2]: Idiap Research Institute – CH-1920, Martigny, Switzerland
[3]: École Polytechnique Fédéral de Lausanne – CH-1015, Lausanne, Switzerland

Abstract

With the increasing number and variety of camera installations, unsupervised methods that learn typical activities have become popular for anomaly detection. In this article, we consider recent methods based on temporal probabilistic models and improve them in multiple ways. Our contributions are the following: (i) we integrate the low level processing and the temporal activity modeling, showing how this feedback improves the overall quality of the captured information, (ii) we show how the same approach can be taken to do hierarchical multi-camera processing, (iii) we use spatial analysis of the anomalies both to perform local anomaly detection and to frame automatically the detected anomalies. We illustrate the approach on both traffic data and videos coming from a metro station.

1. Introduction and Previous Work

An increasing number of camera networks are being deployed to ensure safety and abnormal event detection through visual surveillance. Even if some applications can afford systematic human monitoring, this is surely impossible when the number of cameras in the network is huge. It has become of prime importance to design algorithms able to handle this vast amount of data, filter out typical activities and show the most abnormal parts to human operator. In this article, we improve over recent approaches to do anomaly detection and video abnormality characterization.

Related work A way of detecting abnormality is to learn to recognize abnormal events. From a well specified event type, one can build dedicated detectors [2] that usually perform well. Such approaches require to define the events of interest and gather a variety of training data. These approaches are thus not adequate in large camera networks where no supervision is expected.

Given the above limitations, unsupervised methods have gained interest recently. As these approaches cannot rely on pre-defined abnormality classes, they rather learn what is a normal activity and they consider as an anomaly anything that deviates from these normal activities. Different features have been used to characterize videos. In the context of public spaces, person tracking with person re-identification across cameras provides an effective solution to abnormal behavior detection [12, 1, 13]. However, robust tracking requires sufficient resolution and frame-rate, and, often surveillance cameras have low resolution, low quality (dirty, blurry, etc.) and low framerate (e.g., 5 frames per second). This profile of cameras explain the growing interest in relying on lower level features such as background subtraction information [8] or localized motion in the form of tracklets [5, 6] or optical flow [4].

Probabilistic methods have been shown very effective in handling these low level features in a principled way. Originally designed for text semantic analysis and after their success in many domains, various Topic Models has been proposed and applied for activity modeling [7, 6]. In this article, we build upon the Probabilistic Latent Sequential Motifs (PLSM) model that have been proposed in [9, 11]. The main advantage of PLSM is its capacity of automatically (with no supervision) extracting motifs (temporal patterns) that capture strong temporal information in temporal document represented by $word \times time$ count matrices. Applied to traffic or metro station videos, the motifs are shown to capture the typical activities (related to trajectories) observed in a scene, as illustrated in Fig. 5. PLSM has been used for anomaly detection in surveillance video in [4].

In previous works, PLSM was applied to documents built from an intermediate representation learned by dimensionality reduction of the low-level features. This intermediate representation had been learned in advance which had two drawbacks (i) it made it possible to create artifacts for PLSM, and (ii) the learning ignored temporal information,

and thus was not benefiting by the temporal structure that PLSM can provide. Also, when used for anomaly detection, PLSM was not considering the semantic of the intermediate representation and thus it was for example ignoring the spatial layout of anomaly in the scene. In this article we explore solutions to these two restrictions in PLSM usage.

Outline of contributions We propose to improve the anomaly detection approach using PLSM [9] by (i) jointly learning the dimensionality reduction representation along with the PLSM temporal model, and (ii) reasoning about the spatial distribution of anomalies in the image. Section 2 presents the two main paper contributions. Results are shown in Sec. 3 for traffic data and in Sec. 4 for public space data (metro station). Section 5 concludes the work.

2. Approach and Contributions

The main contributions of this paper are introduced in this section. Subsection 2.2 details how we propose a new model that **integrates** both a **dimensionality reduction** and a higher level **temporal modeling**. The generalization of this extension also allows to hierarchically stack multiple temporal models as also presented in Subsection 2.2 with a multicamera setup. Subsection 2.3 explains how we extract a new anomaly measure that takes into account the **localization of the anomalies**. We start by introducing in Subsection 2.1 the elements that we reuse from state of the art approaches and that motivate our contributions. As the models involve multiple layers, we will systematically use the superscript ll to denote lowest level elements.

2.1. Motif mining using PLSM

Previous approaches that use PLSM extract features from videos and then apply a pre-processing step (e.g., the PLSA topic model in [9]) before PLSM. These steps, introduced below, have been done in a feed-forward manner, the first steps not taking into account the information captured by the subsequent steps. We will show here how this limitation can be waived.

Feature Extraction, low-level words For each video we extract optical flow features from a dense image grid. We keep only pixels where some motion is detected and we quantize the motion into 8 directions, and the 9^{th} "direction" indicates slow moving pixels. We obtain low level words w^{ll} defined by a position in the image and a motion direction. We apply a sliding window of 1 second, without overlap, to obtain an histogram $n^{ll}(w^{ll}, d^{ll}_{t_a})$ indicating the number of times the word w^{ll} appears in the low-level document $d^{ll}_{t_a}$ representing the sliding window at time t_a.

The PLSA model PLSA (Fig. 1, right) is a minimal topic model. Given a set of documents made of word counts and summarized in a table $n^{ll}(w^{ll}, d^{ll})$, it extracts "topics" capturing sets of words that often co-occur in the documents. Each topic is actually a distribution over words $\phi^{ll}_{z^{ll}} = p(w^{ll}|z^{ll})$. Each document d^{ll} is also decomposed as a mixture $\theta^{ll}_{d^{ll}} = p(z^{ll}|d^{ll})$ of these automatically learned topics. PLSA usually has a non-informative prior α^{ll} on the θ^{ll} weight vectors. This means that the model is not encouraged to give any special shape to θ^{ll} distributions and thus it can arbitrarily choose the best ones that explains the data. We will exploit this prior as a mechanism for feeding higher level information to PLSA.

The PLSM model PLSM (Fig. 1, left) adds time to PLSA: it is a topic model which automatically finds temporal and spatial co-occurrences of words. More precisely, it takes as input a count matrix $n(w, t_a, d)$ indicating for each document d (video clips), the number of times the word w occurs at time t_a. By describing the documents as mixtures of temporal motifs, PLSM learns two sets of distributions, similarly to PLSA but adding time. a set of motifs z, each represented by a distribution $\phi_z = p(w, t_r|z)$ denoting the probability that a word w occurs at a relative time t_r since the start of the motif. and the distributions $p(z, t_s|d)$ which indicate when the motifs occur, i.e., the probability that a motif z starts at time t_s. In videos, a motif captures spatially and temporally co-occurring words in the document.

PLSM on top of PLSA One modeling issue with PLSM is how to define the count matrix $n(w, t_a, d)$. A first possibility would consist in ordering the low level documents $d^{ll}_{t_a}$ of the video clip d according to time t_a to obtain the temporal document $n^{ll}(w^{ll}, t_a, d)$. However, as the number of low-level features is quite high, the learning of PLSM can be time-consuming. To overcome this issue, the PLSA topic model can be applied as a dimensionality reduction pre-processing step using as input the un-ordered low-level documents. Figure 1 illustrates this approach with blue arrows. PLSA results in a set of topics z^{ll} which captures the frequently co-occurring words in the video which often correspond to local spatial cluster of words, and the distribution of topics $p(z^{ll}|d^{ll}_{ta})$ within each document. By assimilating these low-level topics z^{ll} as the words w of PLSM, we can build the temporal document for PLSM as:

$$n(w = z^{ll}, ta = d^{ll}, d) = p(z^{ll}|d^{ll}_{ta}) \sum_{w^{ll}} n^{ll}(w^{ll}, d^{ll}_{ta}) \quad (1)$$

where the topic distributions $p(z^{ll}|d^{ll}_{ta})$ are weighed by the mass of the document (number of low-level words at time ta) to account for the overall amount of activity at each time step. This temporal document is then fed as input to PLSM to learn the temporal motifs and their starting times.

Intuition: PLSM as a corrector for PLSA output From the learned distributions ϕ and θ, PLSM is fully able to reconstruct an updated version of its input that takes into account temporal co-occurrence captured in the motifs. The reconstruction of the input is done following the PLSM

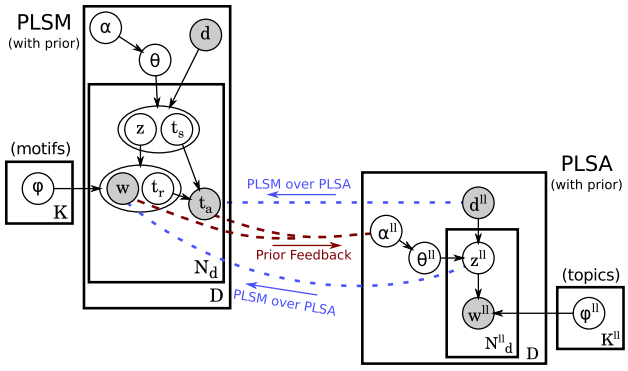

Figure 1. PLSM Model on top of PLSA, from [9] and illustration of the prior feedback (in red). The otherwise uninformative prior on the topic weights in each document $\theta^{ll}_{d^{ll}}$ is replaced by a prior coming from the time-aware higher-level PLSM model.

equations (with $t_r = t_a - t_s$):

$$n^{rec}(w, t_a, d) = \sum_z \sum_{t_s} p(w, t_r|z)p(z, t_s|d) \qquad (2)$$

By applying PLSM on top of PLSA as presented above, PLSM captures temporal patterns of occurrences of PLSA topics. Given that PLSM tries to explain the documents with temporal motifs, the reconstructed input n^{rec} corresponds to the original input but updated to exhibit better temporal coherence. This motif-constrained reconstruction motivated [4] to use the difference between the input and n^{rec} as a measure of temporal anomaly.

We will use the reconstructed PLSM input in two novel ways. First we will consider it as a more time-coherent version of the PLSM input (PLSA output) and feed it back to PLSA in the form of a prior. Next, we will back-project this reconstructed intermediate document n^{rec} down to the input images, to be able to do spatial reasoning about anomalies.

2.2. Integrated hierarchical model

As outlined above, we propose to integrate the PLSA dimensionality reduction step with PLSM. We call this approach Integrated PLSM (IPLSM) and it can be generalized to a hierarchical version for example for multicamera.

Integrated PLSM (IPLSM) To increase the quality of the captured PLSA topics, we propose an integrated model shown in Figure 1 with a red arrow. The goal of IPLSM is to have the temporal structure of the data (captured by PLSM) impacts on the image-level topics (PLSA level). We do it by jointly learning the PLSA and PLSM models. When a scene is crowded, it is often difficult for PLSA to capture clean topics. When getting temporal information from PLSM, PLSA is able to capture cleaner topics. In practice, we exploit two facts: PLSM can reconstruct its input with some added temporal constraints, and, a modified PLSA

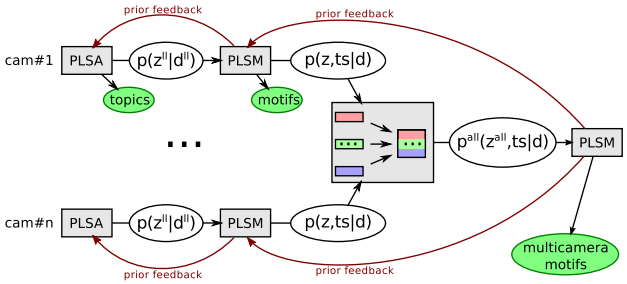

Figure 2. Hierarchical Integrated PLSM Model. Cam#i refers to a camera view i. The process captures information at three levels: topics (image level), motifs (per-camera temporal patterns) and multi-camera motifs. Each level iteratively feeds back information as a prior to the previous level.

can accept a prior on what topics occur in which document. The inter-level information flow is formulated in the form of feedback as explained below.

* Inference by Prior-Feedback The overall goal of IPLSM is to use higher-level temporal structure (motifs) to improve lower level structures (topics). As exact inference is intractable (even just for PLSA), we use iterative approximate inference method and formulate the feedback of higher levels as a prior for lower levels. The document reconstructed by PLSM, n^{rec} will be used after re-weighting as a prior α^{ll} for PLSA. This process goes on iteratively.

More precisely, the prior α^{ll} for the document at time t_a (α^{ll} is used as a shortcut for $\alpha^{ll}_{d^{ll}_{t_a}}$)is computed as follows:

$$\alpha^{ll}(z^{ll}) = n^{rec}(z^{ll}, t_a, d) \times S + Uniform \qquad (3)$$

For a given low level document, α^{ll} represents the hyper parameters of a Dirichlet prior over θ^{ll}. Adding the uniform distributions encodes the inherent uncertainty of the PLSM feedback. This helps in Expectation Maximization algorithm to obtain a better local maximum during the learning phase for PLSA and PLSM. S is a constant and is a parameter to the model: the bigger the S, the more the information coming from PLSM will be trusted. A S value greater than 1 indicates a very strong prior would not allow the motifs to change much as it requires more observations than the training documents to change the belief. So a prior less than one is more meaningful in this context. As previously mentioned, this is an iterative approach, the necessary number of iterations is also a parameter to the model. The number of iterations is camera specific, a metro camera may require more iterations than a traffic camera because of loosely constrained motion. An example of the evolution of topics on a single traffic camera is shown in the result section, in Fig. 5.

Multi-camera with Hierarchical IPLSM Following the idea of integrating PLSA with PLSM, layers of PLSM can be stacked and integrated. The prior feedback can be

used on the α parameter of PLSM. This approach becomes especially interesting when considering multiple cameras as illustrated in Fig. 2. The idea is to have an IPLSM for each camera and a higher-level PLSM working on their combined outputs and capturing motifs of per-camera motifs. The motivation for feedback from hierarchical layer to single camera IPLSM layer is to capture motifs that are relevant (in time) across cameras as depicted in the included supplementary material.

2.3. Localized anomaly detection

As described in Section 2.1, PLSM fits its input with some motifs and it can be used to reconstruct a corrected version of the input. Intuitively, the reconstructed input n^{rec} is the same as the original input when the motifs explains perfectly the input.

Following the above intuition, authors in [4] used the difference between the input and n^{rec} as an anomaly index. One limitation of this approach is that it ignores the semantics of the words used as input of PLSM. In practice, these words correspond to PLSA topics and thus to patches of localized motion in the image.

Low-level document reconstruction Once we obtained $n^{rec}(w, t_a, d)$, as w corresponds to z^{ll} (PLSA topic) it is possible to use the PLSA topics to reconstruct the low level documents. By unrolling the equations, we obtain:

$$n^{rec\,ll}(w^{ll}, d^{ll}_{t_a}) = \sum_{z^{ll}} n^{rec}(z^{ll}, t_a, d) \tag{4}$$

The exact same approach can be used in the hierarchical case with stacked PLSM.

Localized abnormality measure We can obtain abnormalities [3] from the reconstructed document by using the distance measure proposed in [4]. However, this measure does not take into account the spatial locality of the anomaly. We thus compute anomaly by first extracting anomaly in blocks and then finding the most abnormal group of blocks.

We achieve localized abnormality by dividing the frame of video into $h \times w$ sub-frames where h and w are parameters to the model. We compute the reconstruction error measure to each sub-frame as:

$$abn(ta, x, y, d) = \sum_{w^{ll} \in R_{xy}} \left| n^{rec\,ll}(w^{ll}, d^{ll}_{t_a}) - n^{ll}(w^{ll}, d^{ll}_{t_a}) \right| \tag{5}$$

where R_{xy} represents all the words mapping to the sub-frame x, y. We also normalize the reconstruction error in a sub-frame by diving it by the mass of the document corresponding to the sub-frame.

$$normabn(ta, x, y, d) = \frac{abn(ta, x, y, d)}{\sum_{w^{ll} \in R_{xy}} n^{ll}(w^{ll}, d^{ll}_{t_a})} \tag{6}$$

UQM-Junction UQM-Roundabout

Figure 3. The arrows depict driving flow directions which are allowed (in green) or not (in red).

We then use Kadane's algorithm for the **maximum 2D sub-array** problem on $normabn(ta, ., ., d)$ to obtain the abnormality measure for the whole frame and its spatial locality.

3. Experiments on Traffic Videos

To test whether our model can detect activity patterns not learned by the model, we used two different traffic scenes. Below, we will describe in brief the datasets, the anomalies they contained, and provide some quantitative and qualitative evaluation. As parameters for IPLSM, we used 80 PLSA topics, S=0.75, and 4 feedback iterations for learning. The frame segmentation parameter $h \times w$ for the anomaly detector is 24×24. The number of motifs and motif length is specified individually for datasets.

3.1. QMUL Roundabout datset

Dataset It contains 60 minutes of video at a resolution of 360×288 at 25fps. The traffic movements in the roundabout signal are restricted to only certain driving directions as illustrated in Fig. 3. The single type of anomaly present in this dataset is indicated by a red arrow in 3 and corresponds to driving straight ahead on a right only lane. Annotation of these events was conducted on 10 minutes of the dataset.

Parameters and Results The IPLSM model was trained using either 10 or 20 motifs on 15 minutes of video ensured to contain low instances of the abnormality we wanted to detect. As the longest duration for a vehicle to cross the roundabout was around 12 seconds, we choose a motif length of 12 seconds. Examples of extracted motifs are shown in Fig. 5. We applied our method to the test data and compared the results to the ground-truth (a detected event was considered to match the Gt if it overlapped with it). The results are summarized in Table 1.

The false alarm rate was lower when we used higher number of motifs as it could better capture the different traffic patterns variations due to speed, density and type of vehicles in the traffic. We also observed that small vehicles were more difficult to detect in general and would require to set a lower threshold for their detection. Fig. 4 provides

Abnormalities	GT	IPLSM	
		10 motifs	20 motifs
Incorrect direction	15	10	12
False alarms	0	10	4

Table 1. Roundabout abnormality results

Figure 4. Localized anomaly regions detected by our approach. Note that the regions are large as they encompass all the regions with unusual temporal activity (including the regions were activity should have occurred in the normal situation).

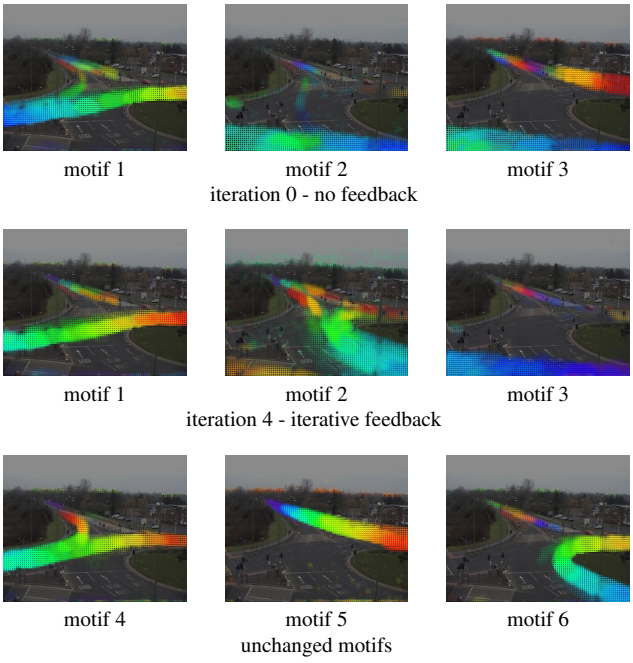

Figure 5. QMUL Roundabout dataset. Example of evolution of motifs during iterative IPLSM learning. Parameters: S=0.75, iteration=5, PLSA topics=80, PLSM motifs=10 with motif length=12. The color gradient represents time from blue (start) to red (12s). Motifs 1-3 evolve during the iterative process from having similar temporal patterns with respect to other motifs in iteration 0 to more distinct patterns in iteration 4. Motifs 4-6 don't evolve as they all represent distinct patterns. Motifs 7-10 are not shown.

some examples of the regions detected by the system.

Abnormalities	GT	IPLSM	
		M=4, ML=80	M=14, ML=10
U-turn	10	7	8
Disruptions	6	4	1

Table 2. Junction abnormality results. M denotes the number of motifs, ML their maximum length.

U-turn disruption

Figure 6. Sample of correct detections (a u-turn and a disruption). Notice that in the disruption case, the vehicles on the left should have closely followed the vehicles on the right so that there are 'missing' cars in the middle.

3.2. QMUL Junction dataset

Dataset This dataset is explained in [6]. In this case, the valid driving trajectories are illustrated in Fig. 3. The abnormalities in this dataset were defined as U-turn and disruptions, where U-turn denotes driving back around the road center, and disruptions indicate interruption of normal flow of traffic by a fire-engine, police or an ambulance. We have not considered Jay-walking as an abnormality because our system is not currently designed to detect abnormalities which reason about the validity of motif occurrences in the context of a cycle Modeling cycles could be done by adding an HMM on top of motifs occurrences [10].

Parameters and Results To evaluate the approach, we trained our approach on 45 minutes of videos (that included the abnormal events) using different parameterization. In one case, we considered motifs of 10s maximum duration, which is more or less the maximum that a vehicle take to cross junction, and of 80s duration, which is the duration of a full traffic cycle. The result are summarized in Table 2. In practice, we observed that traffic disruptions (which often occur out of sync from the traffic cycle) required higher motif lengths able to capture full cycles and provide the necessary context. Also, we noticed that U-turns could not be detected well in very dense traffic, as the generated abnormality were considered negligible as compared to the global activity. Examples of detections are shown in Figure 6.

4. Metro Station Results

This dataset consists of two overlapping cameras, recording neighboring areas: a stairs/escalator(Sc1) and a

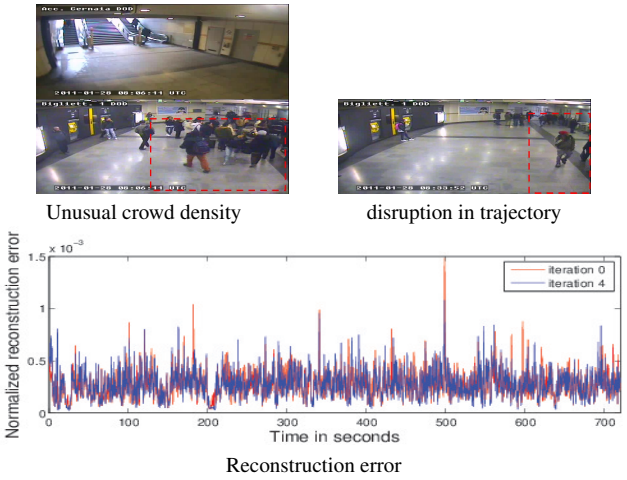

Figure 7. Abnormalities for Metro dataset at 3^{rd} and 2^{nd} level PLSM and effect of feedback on reconstruction error

hall way(Sc2). Sc1 acts as the entrance into the metro station. Sc2 has a ticket counter and turnstiles to the metro network. Below we discuss the effect of feed-back on data reconstruction (more qualitative images in supplementary material) and briefly state the abnormalities detected.

Reconstruction error and abnormalities We trained the hierarchical model with 10 motifs with motif Length of 4 seconds for 2nd level PLSM and 12 motifs with motif Length of 8 seconds for third level PLSM on 720 seconds of video. The rest of the parameters are same as used for the traffic dataset. The motifs learned are not shown here. We performed inference on the trained video using the model obtained at iteration 0 (no feed back) and iteration 4 (with 4 iterations of feedback). We reconstructed the data from both the models from the 3^{rd} level PLSM and obtained a plot of the reconstruction error as shown in Figure 7. As seen from the plot, the reconstruction error is better for iteration 4. We showed in Section 2.2 how the prior feedback improved the sequential patterns obtained. Better motifs should constitute to better reconstruction of data, hence lower reconstruction error. We performed inference on a new video from the same cameras. The abnormalities detected by the system include: unusual density of crowd, people blocking each other and disrupted trajectories. Table 7 shows two of these abnormalities. The abnormality detector for the 3^{rd} level PLSM would detect region in the most abnormal camera while the detectors in the 2^{nd} and 1^{st} level would detect abnormalities pertaining to a single camera.

5. Conclusions and future work

In this paper, we have made two contributions: We formulated IPLSM model which integrates PLSA into PLSM and detect spatially localized anomalies. We also formulated the inter-level information flow as a Dirichlet prior. The feedback mechanism was shown to improve the detected sequential patterns. We then showed how the model can be extended to multiple cameras. We also tested the model on real datasets and showed that it can detect abnormalities and localize the region of abnormality. Presently our model cannot locate multiple spatial abnormalities. Our model also cannot reason about the co-occurrences of multiple motifs. We would like to address these two problems in our future work.

Acknowledgements The authors acknowledge the financial support from the 7^{th} FP of the European Union project VANAHEIM (248907) and from the Swiss National Science Foundation (SNSF) through the PROMOVAR project.

References

[1] N. Anjum and A. Cavallaro. Trajectory association and fusion across partially overlapping cameras. *Advanced Video and Signal Based Surveillance*, 2009. 1

[2] A. Avanzi, F. Bremond, C. Tornieri, and M. Thonnat. Design and assessment of an intelligent activity monitoring platform. *EURASIP Journal on Appl. Signal Proc.*, 2005. 1

[3] S. Calderara, U. Heinemann, A. Prati, R. Cucchiara, and N. Tishby. Detecting anomalies in people's trajectories using spectral graph analysis. *Comput. Vis. Image Underst.*, 115(8):1099–1111, Aug. 2011. 4

[4] R. Emonet, J. Varadarajan, and J.-M. Odobez. Multi-camera Open Space Human Activity Discovery for Anomaly Detection. In *AVSS*, 2011. 1, 3, 4

[5] E. Jouneau and C. Carincotte. Mono versus multi-view tracking-based model for automatic scene activity modeling and anomaly detection. In *AVSS*, 2011. 1

[6] E. Jouneau and C. Carincotte. Particle-based tracking model for automatic anomaly detection. In *ICIP*, 2011. 1, 5

[7] D. Kuettel, M. D. Breitenstein, L. V. Gool, and V. Ferrari. What's going on? discovering spatio-temporal dependencies in dynamic scenes. In *CVPR*, 2010. 1

[8] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *CVPR*, 2009. 1

[9] J. Varadarajan, R. Emonet, and J. Odobez. Probabilistic latent sequential motifs: Discovering temporal activity patterns in video scenes. In *BMVC*, 2010. 1, 2, 3

[10] J. Varadarajan, R. Emonet, and J.-M. Odobez. Bridging the past, present and future: Modeling scene activities from event relationships and global rules. In *CVPR*, 2012. 5

[11] J. Varadarajan, R. Emonet, and J.-M. Odobez. A sequential topic model for mining recurrent activities from long term video logs. *Int. Journal of Computer Vision*, 2012. 1

[12] Y. Wang, L. He, and S. Velipasalar. Real-time distributed tracking with non-overlapping cameras. In *ICIP*, 2010. 1

[13] E. E. Zelniker, S. Gong, and T. Xiang. Global abnormal behaviour detection using a network of CCTV cameras. In *The International Workshop on Visual Surveillance (VS)*, 2008. 1