Multi-camera Open Space Human Activity Discovery for Anomaly Detection

Rémi Emonet[1]
http://idiap.ch/~remonet

Jagannadan Varadarajan[1,2]
vjagann@idiap.ch

Jean-Marc Odobez[1,2]
odobez@idiap.ch

[1]: Idiap Research Institute – CH-1920, Martigny, Switzerland
[2]: École Polytechnique Fédéral de Lausanne – CH-1015, Lausanne, Switzerland

Abstract

We address the discovery of typical activities in video stream contents and its exploitation for estimating the abnormality levels of these streams. Such estimates can be used to select the most interesting cameras to show to a human operator. Our contributions come from the following facets: i) the method is fully unsupervised and learns the activities from long term data; ii) the method is scalable and can efficiently handle the information provided by multiple un-calibrated cameras, jointly learning activities shared by them if it happens to be the case (e.g. when they have overlapping fields of view); iii) unlike previous methods which were mainly applied to structured urban traffic scenes, we show that ours performs well on videos from a metro environment where human activities are only loosely constrained.

1. Introduction and Context

In many visual surveillance set-ups, a human operator has to monitor multiple views. As the number of cameras is huge and the operator can only screen a limited set of views, it becomes critical to design algorithms that automatically pre-select or suggest the cameras to be shown to the human operator. Such pre-selection could be handled by a high level algorithm taking as input some continuous abnormality measure for each camera. In this article, we propose an approach to produce such abnormality measures.

Related work – The first category of approaches for abnormality rating is to explicitly model abnormal events and build detectors for these [3]. These supervised approaches provide strong semantics but require to be able to i) define the events of interest; ii) learn detectors from short amounts of data and iii) handle large variations in view points etc. As ideally, our system should scale to a large number of cameras, we want to *minimize the amount of configuration* like having to define any regions of interest or rules, or gathering and labeling training data on each camera view.

The second class of approaches are the *unsupervised approaches*, as the one we consider in this article. In this category, there have been attempts to model activities in indoor scenes like a metro station scene. To handle multiple cameras, approaches generally rely on predefined calibration information between cameras as in [10] or learn it automatically [11]. However, the common problem of most existing multi-camera approaches for human activity analysis is that they rely on person tracking [12, 2, 11] or re-identification. In our context, we process poor quality videos with a lot of occlusions and tracking in such conditions is still a research challenge, so we prefer *lower level features* such as instantaneous motion or background subtraction as in [6].

In parallel, with the success of *topic models* in unsupervised learning in different domains, advanced models were proposed for activity modeling such as in [5, 4]. This class of approaches have the ability to discover dominant activity patterns occurring in the scene using simple low-level features. In this article, we use one such model called *Probabilistic Latent Sequential Motifs* (PLSM) [9]. PLSM, unlike other topic models, represents activities as temporal patterns called *motifs*. These motifs typically correspond to dominant activities in the scene. Furthermore, they also enable us to infer their start times in a new test video.

Existing topic models were demonstrated only on highly constrained scenes such as traffic scenes and with a single viewpoint [5, 9, 4]. In our case, we explore the use of these models in the context of a metro station that contains *multiple cameras and loosely constrained activities*.

Article Structure – Section 2 introduces our approach that uses a state of the art temporal topic model to extract normal activities. We also explain how we produce an abnormality measure from these activities. In section 3, we validate the model ability to capture meaningful activity patterns using synthesized video data. In section 4, we apply the model on real videos from a metro station and conclude in section 5.

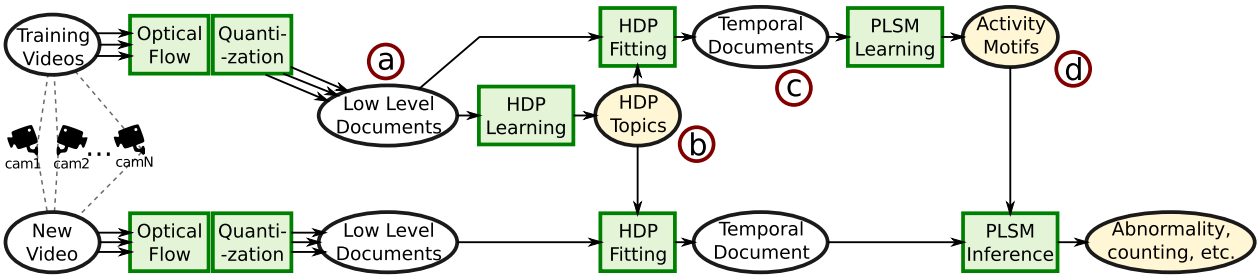

Figure 1. Overall process used: low level feature extraction and quantization, dimensionality reduction using HDP, temporal activity extraction using PLSM. (HDP: Hierarchical Dirichlet Process, PLSM: Probabilistic Latent Sequential Motifs)

2. Proposed Approach

2.1. Overview of the Approach

Our approach towards abnormality detection is to first learn normal activities in the scene. Then, we consider as abnormal any activity that cannot be sufficiently explained as a mixture of these learned normal activities. An abnormality can be roughly categorized into two types [7]: 1) events that are fundamentally unusual in appearance, and 2) unusual order of events, where all or most of these events are normal. The adopted PLSM model [9], due to its complete representation of scene activities, can simultaneously capture both kinds of abnormalities.

Compared to [9], we use different low level features and a different low-level processing. To avoid the need for a complex background model, we only use quantized optical flow features. We further process these features with a *Hierarchical Dirichlet Process* (HDP) [8] (in place of Probabilistic Latent Semantic Analysis, PLSA). HDP has the advantage that it automatically finds the number of topics that best matches the observed data.

In the next subsection, we introduce notations and explain how we create the "temporal documents" used as input of PLSM. We also explain how we use the output of PLSM to do rate the abnormality of a scene.

2.2. Two-Level Topic Models

As introduced in section 2.1, we use two levels of topic modeling: first HDP, then PLSM. Fig. 1 depicts the overall process, including both training and runtime phases. To avoid confusion in the notation, we will systematically use the "ll" superscript for low level (HDP) elements.

Feature extraction – For each view, we first extract optical flow features (motion direction) on a dense image grid. We keep only pixels where some motion is detected and for these, we quantize the motion into 9 "categories": one for each of the 8 uniformly quantized directions and one for a "really-slow" motion. A low-level word w^{ll} is defined by a camera index, a position in the image and a motion "category". The size of this low level vocabulary is usually around 30000 words (considering only words that are actually observed). On these low-level words, we run a sliding window of 1 second long (5 frames), without overlap. For each second ta, we obtain an histogram $n^{ll}(w^{ll}, d^{ll}_{t_a})$ of low-level words from all cameras (Fig. 1a) in the corresponding window. Here, $d^{ll}_{t_a}$ is the low level document obtained from the sliding window at a time ta.

Multi-Camera – As illustrated in Fig. 1, the camera views are fused right after feature extraction. The subsequent processing, and particularly HDP and PLSM, perform co-occurrence analysis that automatically find and exploits possible relations between cameras (more in section 4.1).

Low-level HDP – On the (unordered) set of documents $\{d^{ll}_{t_a}\}_{ta}$, we apply the HDP topic model, for which a publicly available implementation was used [1]. The goal is to perform a dimensionality reduction: HDP learning takes as input $n^{ll}(w^{ll}, d^{ll}_{t_a})$ (word counts for the each $d^{ll}_{t_a}$) and outputs as set of "topics" (Fig. 1b). Each topic z^{ll} is defined as distribution $p(w^{ll}|z^{ll})$ over the words and corresponds to a soft cluster of words that regularly co-occur in documents.

HDP (both learning and fitting) also outputs another information that is, for each document $d^{ll}_{t_a}$, a decomposition of it as a mixture of existing topics, expressed as the distribution $p(z^{ll}|d^{ll}_{t_a})$. We use this information, re-weighted by the amount of activity at instant ta, to build the temporal documents (Fig.1c) that will be the input of PLSM. A temporal document d is expressed as $n(w, ta, d)$, the (high-level) word counts at each time ta instant in the temporal document. Here is the corresponding formula:

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

Higher-level temporal motifs (PLSM) – Given the temporal documents obtained after HDP, we directly apply the PLSM algorithm [9]. PLSM takes as input a set of temporal documents $n(w, ta, d)$ and decomposes it as a set of temporal motifs (z) and when they start in each document (ts). More precisely, PLSM "un-mixes" the documents by learning jointly the recurring motifs (Fig.1d) and their starting times. Each motif z is defined by a probability distribution

$p(w, tr|z)$ indicating the probability of observing the word w after tr time steps from the beginning of the motif. For each temporal document d, the starting times are described by a distribution $p(ts, z|d)$ over the variables ts (starting times) and z (motifs).

Testing on new videos – For new videos of the same viewpoint, we can create the corresponding temporal document using the same process as during the learning, keeping the HDP topics fixed. In the same way, we then process the obtained temporal document with PLSM, keeping the activities (motifs) fixed. Eventually, we have the motifs (from the learning phase) and their starting times $p(ts, z|d)$ for the temporal documents corresponding to the new videos.

2.3. Measure for Abnormality Rating

We can extract an abnormality measure from the motifs and their optimal starting times $p(ts, z|d)$. The measure we use is a "reconstruction error" defined as the distance between the original document and the document reconstructed as a mixture of the motifs $p(w, tr|z)$ occurring at $p(ts, z|d)$. In fact, the observed temporal document is a matrix of counts $n(w, ta, d)$ but the PLSM model produces a probability table $p(w, ta|d)$ of a word w occurring at the absolute time ta:

$$p(w, ta|d) = \sum_{ts} \sum_{z} p(ts, z|d) p(w, tr = ta - ts|z) \quad (2)$$

To allow direct comparison, we normalize the whole temporal document (so that it becomes a probability table). Then, our abnormality measure at a given instant ta is given by:

$$abnorm(ta, d) = \sum_{w} \left| \frac{n(w, ta, d)}{n(d)} - p(w, ta|d) \right| \quad (3)$$

Where $n(d)$ is the total number of observations in the temporal document d.

3. Model Behavior on Synthetic Activities

Virtual Scene To validate the implementation of PLSM and to check if it recovers meaningful motifs, we generate synthetic "videos" featuring red circles moving along some predefined trajectories. Fig. 2 depicts the single-view synthetic scene with the different trajectories. On these videos, apart from some minor artifacts due to the absence of texture, the optical flow extraction is perfect.

Each person (red dot) follows either the T1, T2 or T3 path, at a constant but randomized speed. A person on T3, will take the T3A route by default but will take the T3B one if another person is arriving on T2 (and is close enough).

Recovered HDP Topics and Activities After running HDP (see Fig. 1), we obtain some topics that represents localized activities in the image. From a model perspective,

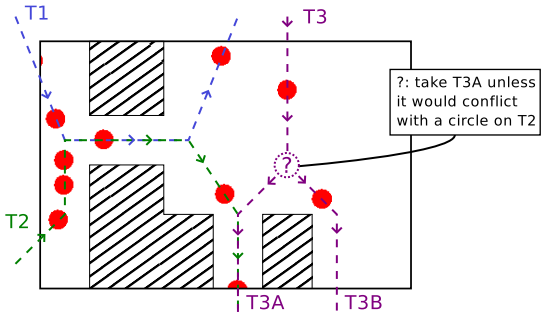

Figure 2. Scene used for synthetic experiments (see section 3).

Figure 3. Four topics distributions $p(w^{ll}|z^{ll})$ among the 31 extracted. For improved readability, we represent a topic $p(w^{ll}|z^{ll})$ by only considering the location of the words w^{ll} (we do not represent motion direction). Other topics are like these four, covering a part of the trajectories in the scene.

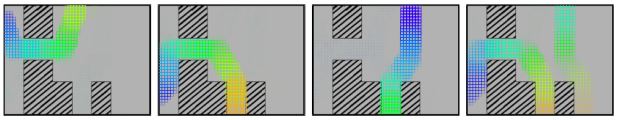

Figure 4. The 4 recovered motifs ($p(w, tr|z)$ for the 4 z values) corresponding to T1, T2, T3A and T2+T3B activities. Color gradient represents time (tr), starting from blue.

we obtain a number of topics and, for each of them, a distribution over the words $p(w^{ll}|z^{ll})$. With the scene presented above, we obtain 31 topics that properly segment the existing motion patterns. Fig. 3 shows examples of four topics taken among them.

These 31 HDP topics are used to build the temporal documents as explained in section 2.2. We then use PLSM to extract motifs from these temporal documents. In Fig. 4, we show the motifs obtained when asking PLSM to find 4 motifs. We see that we recover the four activities T1, T2, T3A and T3B. It is interesting to note that T3B never happens without a synchronized occurrence of T2: the motif captures this systematic co-occurrence.

4. Abnormality Rating in a Metro Station

We are interested in modeling normal activities in videos taken from cameras in a metro station. In this section, we apply the same procedure as in section 3 but on real video data. We first show some retrieved activity motifs obtained from PLSM on the videos and then go on to present how they are used for abnormality detection.

4.1. Retrieved Topics and Motifs

Scenes and Parameters – To test our model under different conditions, we considered two different scenes. The *first scene* (Sc1) is made of two cameras, with non-overlapping viewpoints, recording two neighboring areas: a stairs/escalator area and a walking area. The *second scene* (Sc2) is made of the hall of a metro station which is recorded by two cameras with overlapping viewpoints. The hall is connected to the two station entrances, contains two vending machines and has a row of turnstiles used as entry or exit points to the metro network.

Used Parameters – We used two hours of training video in 704x288 at 5 frames per seconds, for each camera of each scene. In the experiments shown here, HDP selected a number of low-level topics between 70 and 80 (depending on the runs). Also, we configured PLSM with a maximum motif length of 15 seconds and asked for 20 motifs.

Multi-camera and HDP – Each scene is a video montage of two cameras that is then processed as in Fig. 1. The features computed therefore come from both the views. The HDP topics (not shown for space reasons) properly capture multi-camera relationships in spite of the almost continuous people activity. HDP topics spans the two views when there is an overlap between them (as in Sc2) and are limited to individual views when there is no such overlap (as in Sc1).

PLSM motifs on Sc1 – Fig. 5 illustrates the motifs we recover for Sc1. These motifs are highly meaningful:

- We nicely recover the person going up using the escalator in a) and down using the stairway in b).
- On the upper area, we recover the different trajectories. For example, we get the exit trajectories coming either from the (non visible) mezzanine in c) or from the stairway in d) (with the motif correctly spanning over the two cameras).
- We also get entering trajectories (from the turnstiles). We get multiple variations (6 of them) that cover different positions (persons more or less close to the camera) and different speeds. One speed variation is shown in e) and f) which differ only by the color palette: f) ends in green meaning it is slower/longer than e).

PLSM motifs on Sc2 – Fig. 6 shows 9 representative motifs retrieved for Sc2. Here again, we obtain motifs that represent classes of activities:

- In a) and b), we recover people entering the station and going to the nearest turnstiles.
- In c), we can observe that people entering from the left side often go to the right side. This behavior is explained by the presence, on the right, of an elevator and the escalator for going down to the platform.
- In d), we capture people coming from the escalator (blue in the middle left part of the lower image), and then exiting the station (in red) after passing behind the pillar. We also see that this behavior is actually correlated with some activity on the right: we see here the two ways of reaching the station exits from the metro platform.
- In e) and f), we see the typical trajectory of people that just use the station as an underground passage without actually going beyond the turnstiles.
- In g), h) and i), we show one of the ways of going to the vending machines in g); and the two ways of leaving the vending machines (h) for the one on the left, i) for the one on the right).

4.2. Extracted Abnormalities

Given the strong semantic of the motifs extracted for the views in the metro station, we try to use them for abnormality rating and expect meaningful results. We evaluate the abnormality measure (reconstruction error) as presented in section 2.3. Scene Sc1 is more constrained, so we rather illustrate the results on Sc2. Fig. 7 shows an annotated plot of the abnormality measure for Sc2:

- The high abnormality in the beginning, cf a), is explained by a group of 7 persons moving from place to place in the hall, mostly with non straight movement.
- Abrupt changes in trajectory, such as in b), also cause an increase in abnormality. In such a case, the temporal aspect plays an important role: each subpart of the trajectory is mostly normal.
- Numerous abnormality peaks are caused by people blocking each other. This is the case for all boxes in green. In the typical example of c), three groups have conflicting trajectories and this will cause most of the people to slow down and change trajectory.
- In d), the peak is caused by tourists with rolling suitcases that first stop then move along turnstiles.
- The case of e) is a surprisingly rare event: a young woman arrives running in a curved trajectory, then falls down and is joined in a hurry by some friends.

5. Conclusion

In this article, we presented an approach for abnormality rating of video streams. It consists of two steps, first, automatically finding recurrent (normal) motion activity patterns, and second, measuring abnormality as the deviation from the learnt normality. The normal recurrent activity patterns are extracted using two levels of topics models. First, Hierarchical Dirichlet Process (HDP) is used at lower level to operate a dimensionality reduction by capturing the groups of co-occurring low-level motion features. Second,

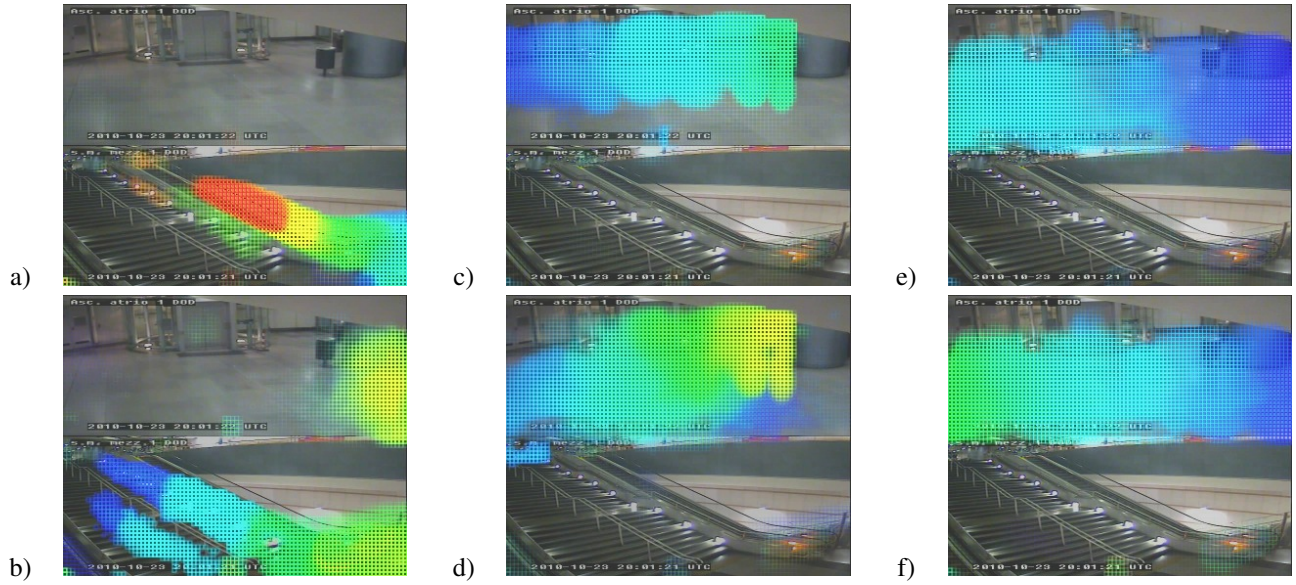

Figure 5. *[best viewed in color]* Representative motifs for Sc1 (for comments, see the body of the article). Color gradient represents time with blue: start of the motif, green: middle (7 seconds later), red: end of the motif (14 seconds later).

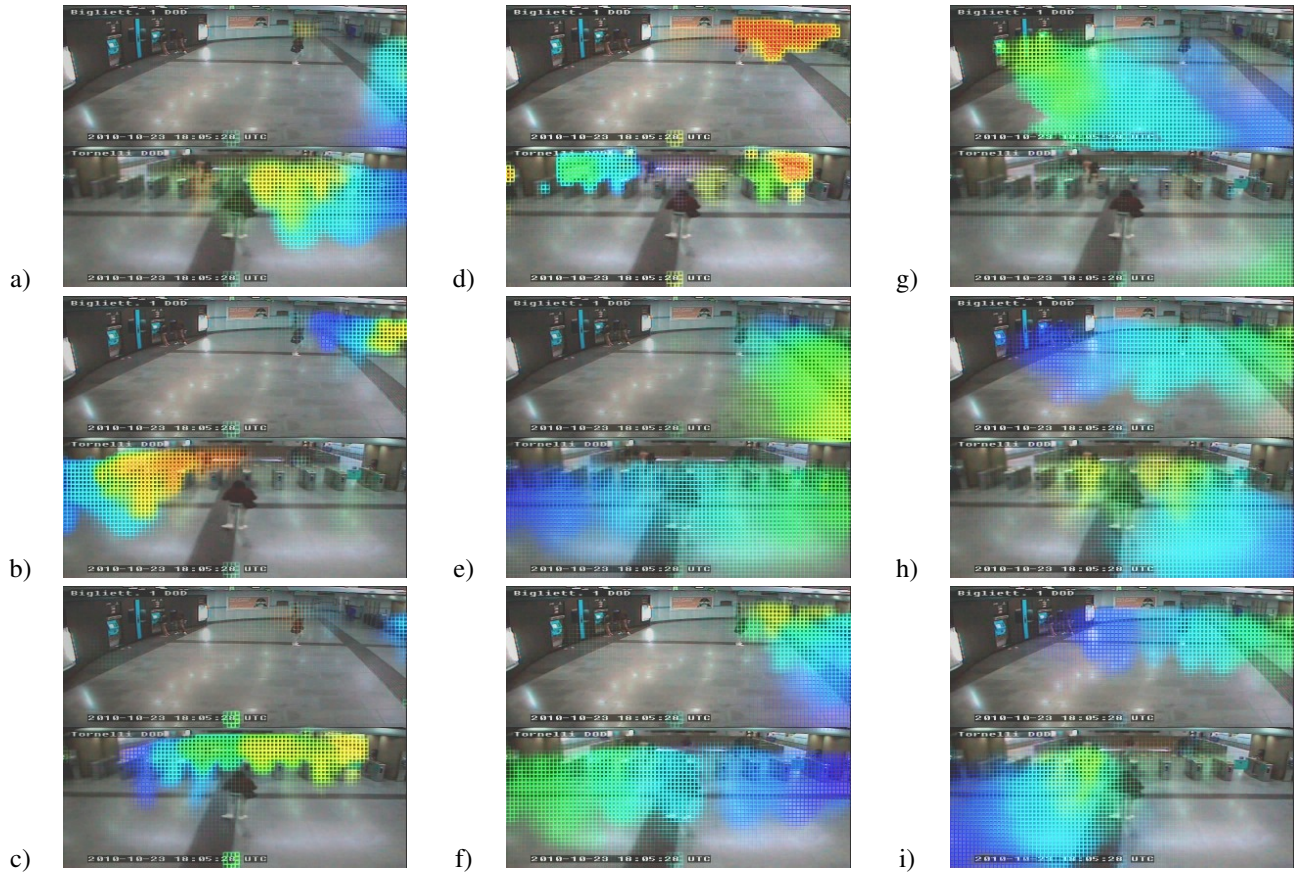

Figure 6. *[best viewed in color]* Representative motifs for Sc2 (for comments, see the body of the article). Color gradient represents time with blue: start of the motif, green: middle (7 seconds later), red: end of the motif (14 seconds later).

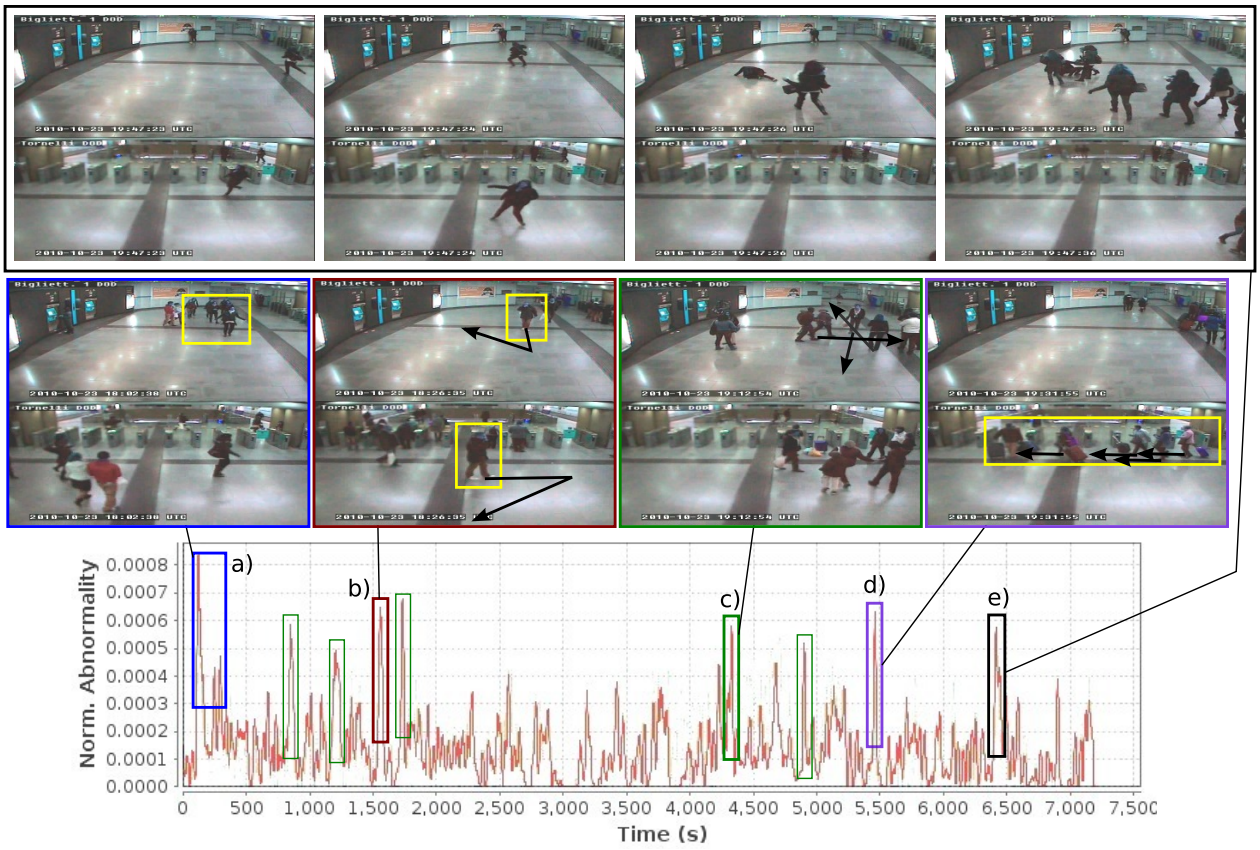

Figure 7. Abnormality detection on Sc2 using PLSM motifs

at a higher level, we modeled temporal information using Probabilistic Latent Sequential Motifs (PLSM).

We obtained meaningful motifs from PLSM, each corresponding to a typical human activity. From these motifs we showed that we were able to extract interesting events from a video of 2 hours. We showed that such an approach can be applied to loosely constrained scenes such as human motion in a metro station. In particular, we demonstrated that the method can jointly process and correctly handle multiple cameras (without any calibration), enabling to monitor automatically larger areas in the metro station.

Acknowledgements – The authors gratefully acknowledge the financial support from the Swiss National Science Foundation (Project: FNS-198,HAI) and from the 7th framework program of the European Union project VANAHEIM (248907).

References

[1] `http://www.cs.princeton.edu/~chongw/`.

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

[3] 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.

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

[5] 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.

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

[7] P.Cui, L. Sun, Z.Q.Liu, and S.Yang. A sequential montecarlo appraoch to anomaly detection in tracking visual events. *IEEE Workshop on Visual Surveillance*, pages 1–8, 2007.

[8] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.

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

[10] X. Wang, K. Tieu, and W. E. L. Grimson. Correspondence-free multi-camera activity analysis and scene modeling. In *CVPR*, 2008.

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

[12] 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.