IDIAP RESEARCH REPORT

UNSUPERVISED METHODS FOR ACTIVITY ANALYSIS AND DETECTION OF ABNORMAL EVENTS

Remi Emonet Jean-Marc Odobez

Idiap-RR-21-2013

MAY 2013

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Chapter 13

Unsupervised methods for activity analysis and detection of abnormal events

13.1. Introduction

The use of video-surveillance cameras serves numerous requirements and various objectives:

- an objective of safety, when it is a question of ensuring the physical safety of people in a certain environment – e.g. when passengers are boarding or getting off a train or subway car, or to detect incidents which could cause accidents on a freeway or in an urban environment;
- an objective of security and protection of equipment, by detection of intrusions, of unattended luggage, acts of aggression and generally any antisocial behavior or vandalism;
- an objective of efficiency, by identifying tendencies in flows so as to detect blockages (traffic jams in a road situation for instance) or prevent them by the appropriate means – information and recommendations to the users, modification of itineraries, etc.

Yet, in the vast majority of cases, the cameras are simple recording boxes, whose data are only exploited *a posteriori* when a crime has been committed. This is because of prevailing sentiments or legal reasons (to safeguard people's privacy), but also because of technical and economic factors: the automatic analysis algorithms are not sufficiently reliable, or the infrastructural costs relating to the use

---

Chapter written by Rémi EMONET and Jean-Marc ODOBEZ.

of such algorithms or to the employment of surveillance operators are too high in view of the large number of cameras needing to be analyzed. Besides their performance, a significant factor limiting the use of algorithms relates to configuring them when deployed in a real-world environment: this is generally a technical and difficult task, performed by non-specialists in the domain – that is to say, people who are not conversant with the issues relating to computer vision.

Against this backdrop, a considerable amount of effort is currently being investigated in creating algorithms which, based on observations of a time period between an hour and several days, are capable of deducing the typical activities in a scene; when those activities begin and end; the relations between them; the times when they are most likely to occur; and so on. Such information may be useful in itself, in order to better understand the content of the scene and its dynamics, or in pre-treatment prior to higher-level analysis. For instance, analysis could help discover the actual activities from the sensor's viewpoint, provide a context for other tasks (such as tracking or interpretation of the data) or define indicators of abnormal situations which can be exploited to automatically select the streams to display to an operator in a control room monitoring hundreds of cameras.

In this chapter, we present a recent group of approaches oriented in this direction. Based on so-called topic or theme models, in reference to the context in which they were originally developed – that of semantic analysis of texts – these unsupervised (or at least, not heavily supervised) approaches are able to discover the main activities in a scene, possible cycles, anomalies, etc. by analyzing the co-occurrences of visual words. These visual words are defined by quantifying simple characteristics of the video – such as the position in the frame (and by extension, in the real-world scene), the apparent motion, indicators of size and shape, etc. – which are extracted immediately, thereby avoiding having to track the objects in the scene: a task which is currently tricky in practice for crowded environments.

The chapter is organized as follows. In the first part, we begin by studying the main concepts of topic models through the lens of one of the simplest such models: Probabilistic Latent Semantic Analysis (PLSA). We shall then show, in the same part, how this method can be applied to activity seeking in videos by defining a vocabulary (choice of words and what they represent) and appropriate documents. In the second part, we present a more recent model of our own design, and which is able to discover temporal topics (hereafter called *motif*s), i.e. topics which do not simply capture the co-occurrence of words at a given time as PLSA does, but also the order in which those words occur over the course of time. Both these parts are illustrated by results which visualize the activities discovered. In the third part, we give examples of the use of these models – e.g. for anomaly detection or prediction – and suggest a number of ways in which they might be evaluated. The chapter

closes with a discussion of projects currently in progress and future work to be envisaged in this domain.

13.2. An example of a topic model: PLSA

13.2.1. *Introduction*

Recently, the creation of Bayesian probabilistic models called topic models has become an avenue of research, which is relevant for discovering recurring patterns in data provided by all kinds of sensors. These models originate in the domain of automated text analysis. They consider a text as a bag of words (BOW), obtained by counting the number of occurrences of each word in the document, thereby eliminating all information about the order in which they appear. In spite of this simplification, because the words contain a substantial amount of semantic information, BOWs are a representation which has been successfully used for numerous tasks of textual analysis, such as classification into different genres or text retrieval. Topic models, such as PLSA [HOF 01] or the LDA (Latent Dirichlet Allocation [BLE 03]) model, are built around BOWs, and have been introduced to discover the prevailing topics in datasets by analyzing the co-occurrence of the words: a notion similar to correlation but which applies to discrete data.

Because of their analytical power, their easy implementation, their versatility and their unsupervised nature, topic models have been applied to a great many problems and modalities as a data-mining tool. In particular, they have been used in different forms to discover human activities in sport videos [NIE 08], surveillance footage [WAN 09], accelerometer data [HUY 08] or GPS coordinates from mobile telephones [FAR 08]. However, the specification of a vocabulary and of documents appropriate for the discovery of activities of interest, the actual modeling of spatial and temporal information, the interpretation of results and the detection of abnormal events still represent considerable challenges, both in general and for a specific domain of application.

In this section, we shall give a more detailed presentation of the PLSA model, and then explain how it can be applied to videos to discover activities.

13.2.2. *The PLSA model*

The PLSA model [HOF 01] was introduced as a probabilistic version of latent semantic analysis (LSA) to capture recurrent co-occurring information in a discrete dataset. Although it is considered a not-entirely generative model, the simplicity of

its optimization procedure makes it an interesting alternative to entirely generative models such as LDA [BLE 03].

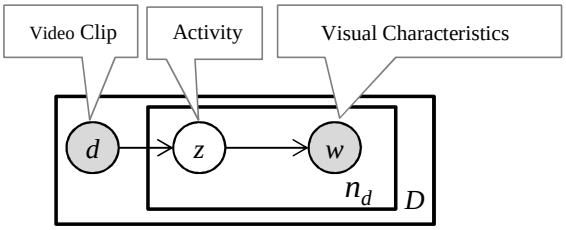

Figure 13.1. *The PLSA generative model. The shaded nodes represent observed variables, whereas the other node denotes latent variables. The rectangles indicate an identical repetition of the process described by its content*

13.2.2.1. Generative process

The graphic model is given in Figure 13.1. For a set of documents *d*, the process of generation of the observations (the (*w*,*d*) pairs of words appearing in the documents and forming the BOWs) is as follows:

- randomly draw the document *d* (in which the observation will be generated) according to the probability $p(d)$;
- draw the topic $z \sim p(z|d)$, where $p(z|d)$ represents the probability of finding the topic *z* in the document *d* – that is, indirectly, the probability that a word *w* in the document will belong to the topic *z*;
- draw the word $w \sim p(w|z)$, where $p(w|z)$ is the probability that the word *w* will appear in the topic *z*.

As this process shows, PLSA assigns each observation (*w*, *d*) an associated latent variable $z \in Z = \{z_1, ..., z_{Nz}\}$ defining the topic of the observed word. The joint probability of the resulting model is then given by:

$$p(w,z,d) = p(d) \,.\, p(z|d) \,.\, p(w|z) \qquad [13.1]$$

From a probabilistic point of view, this introduces a hypothesis of conditional independence between the variables observed, i.e. that the appearance of a word is independent of the document, given the topic to which it belongs. With this model, the likelihood of observations is therefore:

$$p(w,d) = p(d) \cdot p(w|d) = p(d) \cdot \sum_{z=z_1}^{z_{N_z}} p(z|d) \cdot p(w|z) \qquad [13.2]$$

As this expression indicates, the model decomposes the distribution $p(w|d)$ of the words in a document into a convex linear combination of the topics $p(w|z)$, where the weights $p(z|d)$ are given by the distribution of topics in the document. We then have a typical mixture model, just like mixtures of Gaussians for continuous data.

13.2.2.2. Inference

The estimation of the parameters Θ (in our case the different probability tables, which gives us $\Theta = \{p(d), p(z|d), p(w|z)\}$) is typically done using the principle of maximum likelihood. More specifically, given a training dataset D, the log-likelihood of Θ is expressed by:

$$L(\Theta|D) = \sum_{d \in D} \sum_{w} n(d,w) \cdot \log(p(w,d)) \qquad [13.3]$$

where $p(w,d)$ is given by equation [13.2]. In practice, in view of the presence of the sum in the logarithm, optimization is performed using a standard iterative EM (Expectation-Maximization) algorithm whereby the probabilities of the hidden variables are estimated and then used in a stage of maximization of the parameters [HOF 01]. This procedure leads to the estimation of the topics $p(w|z)$ and the distributions of the topics $p(z|d)$ in the learning documents.

13.2.2.3. New documents

In this case, we are interested only in estimating the weights $p(z|d)$ of the topics in the new document d. These are obtained using the same EM algorithm as above, but without updating the topics $p(w|z)$, and which simply leads to the maximization of the normalized log-likelihood L^{norm} in each document d:

$$L_d^{norm}(p(z|d)) = \frac{1}{n_d} \sum_{w} n(d,w) \cdot \log\left(\sum_{z} p(z|d) \cdot p(w|z) \right)$$
$$\text{avec}: n_d = \frac{1}{n_d} \sum_{w} n(d,w) \qquad [13.4]$$

13.2.3. *PLSA applied to videos*

The PLSA model can be applied to any type of data. In video analytics, we want the topics discovered to characterize frequent activities in the scene. In practice, the semantics of the topics will depend essentially on the definition of the vocabulary and the way in which the documents are constructed. In this section, we present a

simple example of construction of a vocabulary and documents, and illustrate the results obtained.

13.2.3.1. Vocabulary

The vocabulary must characterize the content of the scene, and is obtained by quantifying simple characteristics in the video. The typical example in the existing body of literature is based on two characteristics: position and motion:

- position: in surveillance videos, the activities are often characteristics of the place where they are occurring. It is therefore helpful to take account of position when constructing the vocabulary, e.g. by quantifying the position into cells (or blocks) of 4×4 to 10×10 pixels;
- motion: motion is an essential piece of information in order to differentiate activities. The estimation can be performed robustly (e.g. using the multi-resolution Lucas-Kanade algorithm) as regards variations in content (texture), and at a reasonable computation cost. It is particularly advantageous for our purposes because it is relatively independent of the lighting conditions. In order to be used as a word, the motion must be quantified. Conventionally, the direction is deemed the most important factor, and sufficiently great motions are classified by quantifying their direction into four or eight labels.

We can then define the vocabulary as the Cartesian product of the position and motion spaces. Thus, for a 280×360-pixel image, and using 4×4-pixel blocks, we get a total of $70 \times 90 \times 8 = 50{,}400$ potential words. In practice, during learning, this set can be reduced by eliminating all the words which never appear or which represent less than 0.5% of observations. Finally, we usually get around 10,000-20,000 words.

13.2.3.2. Documents

These are simply constructed by dividing the video into short clips. We thereby obtain the BOW for each document d by counting the number of times $n(d,w)$ that it contains each word w.

13.2.3.3. Example of topics discovered

The PLSA method can be illustrated by considering a 1h45m video of the scene visible in Figure 13.3. In this case, the activity of a vehicle may be described as a set of movements (position and direction) which co-occur in the clip. Each activity thus corresponds to a topic represented by the distribution $p(w|z)$ of the visual characteristics which frequently co-occur.

In order to identify the image positions where the activities associated with a given topic occur, for every position c, we can marginalize the topic distribution over to the set of words V_c attached to that position (and with different directions of movement). We thus obtain an activity map:

Figure 13.2 illustrates this marginalization on the scale of the whole image.

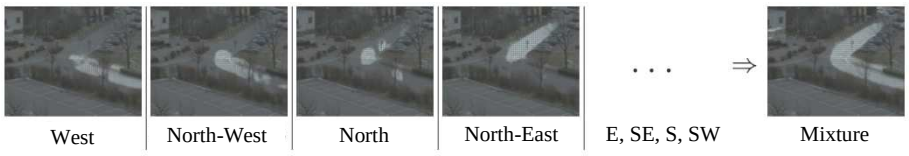

Figure 13.2. *Re-projection of a topic (twenty-second window) on each of its eight directions, and then in a condensed version combining the eight directions. The four directions omitted from this sequence (E, SE, S, SW) do not contain any activity in this example*

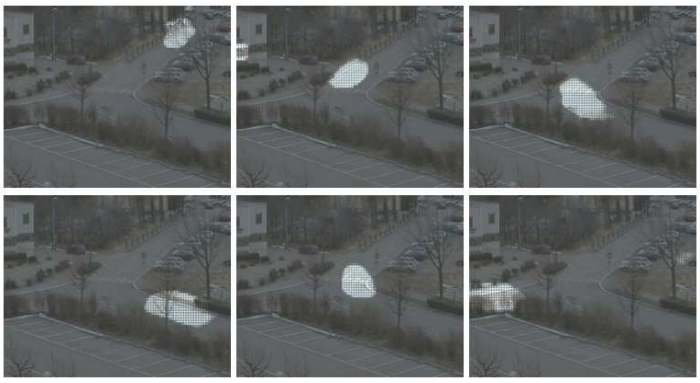

Figure 13.3. *Examples of topics obtained (6 out of 75) using one-second video clips as a document*

The images in Figure 13.3 show a few topics found when the duration of the clips is one second and we wish to find $N_z = 75$ topics. These correspond to the elementary activities which can happen in the space of a second, and enable us to reconstruct the total activity observed over a certain amount of time by linear combination.

In order to capture more semantic activities, we can increase the duration of the clips. The results with a window of ten seconds (and $N_z = 20$) are presented in Figure 13.4. Although the activities are captured in this example, note that the time-

related information is lost, so it will not be easy to determine the exact moment at which a certain activity takes place, e.g, in cases where multiple cars are following each other.

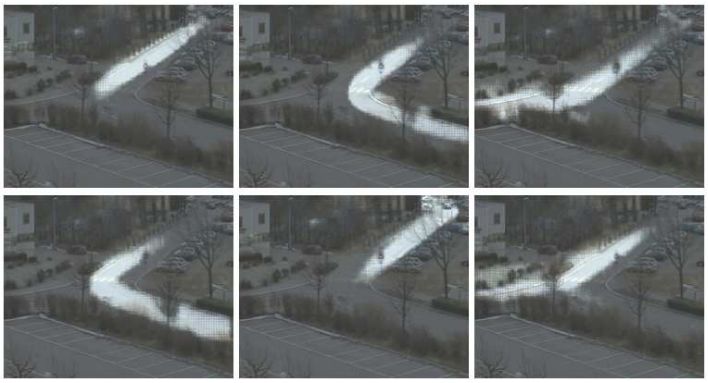

Figure 13.4. *A few topics obtained (six out of twenty) using ten-second video clips as a document. PLSA indeed enables us to find the main activities in the scene. It should be noted that the temporal information is lost*

13.2.3.4. Influence of the vocabulary

The previous examples emphasized the influence of the duration on the topics found. By visualizing the set of results, we could also show the absence of topics that represent cars stopping at the crossroads. Indeed, there are no words which capture such information. In [VAR 09], this is taken into account by applying a background subtraction algorithm. This allows us to create words with the label "static": these represent points in the foreground whose estimated motion is zero. In that case, after applying PLSA, we find specific topics related to that characteristic. Note that in the same article, the use of words related to the size of the blobs obtained by background subtraction also enables us to clearly distinguish (in a different scene) the activities of pedestrians from those of cars, particularly at pedestrian crossings.

13.3. PLSM and temporal models

As explained in the previous section, simple topic models like PLSA enable us to capture the recurrent activities in a scene but in doing so lose all temporal information contained in a document. The PLSM model (for Probabilistic Latent Sequential Motifs) enables us to get around this problem and capture temporal information in a topic which we then call a "motif". Other approaches try to model time but, unlike PLSM, are incapable of separating the recurrent activities and

finding when they appear as well. This section is devoted to a presentation of the PLSM model.

13.3.1. *PLSM model*

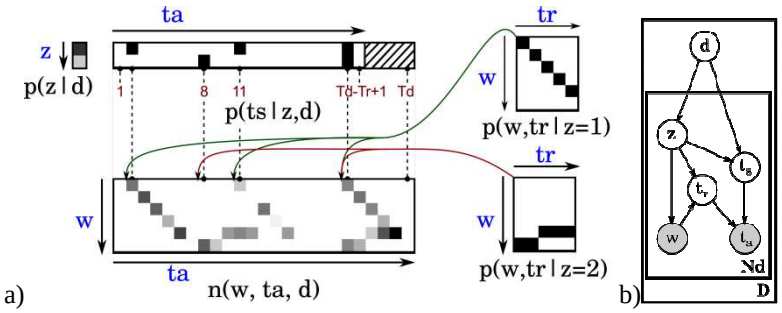

Figure 13.5. PLSM *Generative Process: a) generation of a temporal document d; b) graphical model (the observations are shaded)*

As input, the PLSM model takes a set of temporal documents defined by a count matrix $n(w,t_a,d)$ formed of a collection of accumulated observations, each observation being a (w,t_a,d) triplet. As illustrated in Figure 5a, PLSM models a temporal document as a combination of latent elements:

- motifs (two in the example): each motif z is a distribution $p(w,t_r|z)$ defined on the Cartesian product of the vocabulary and a temporal axis;
- the moments in the temporal document where the motifs appear: each motif z begins according to a table $p(t_s|z,d)$.

13.3.1.1. Generative process

The graphic model of PLSM is given in Figure 13.5b. For a set of temporal documents, the process of generation of each observation (w,t_a,d) is as follows:

- randomly draw document d in which the observation will be generated;
- draw the motif: $z \sim p(z|d)$ where $p(z|d)$ is the probability that an observation from the document d will come from a motif z;
- draw a time of occurrence: $t_s \sim p(t_s|z,d)$, where $p(t_s|z,d)$ is the probability of a motif z beginning at time t_s in the document d;
- draw a word and a relative time: $(w,t_r) \sim p(w,t_r|z)$[1], where $p(w,t_r|z)$ is the probability that a word w will appear at a time t_r in a motif z;

[1] In Figure 14.5b, the selection of (w,t_r) is broken down into first drawing the relative time t_r followed by drawing the word w given the knowledge of t_r.

- assign $t_a = t_s + t_r$ (t_a is completely defined, when t_s and t_r are known).

The joint distribution on all the variables of the model can be derived from the generative process. Given the deterministic relation $t_a = t_s + t_r$, only two of these three variables generally appear in the equations. The joint distribution for the PLSM model is as follows:

$$p(w,t_a,d,z,t_S,t_r) = p(d)\cdot p(z|d)\cdot p(t_S|z,d)\cdot p(w,t_a - t_S|z) \quad [13.5]$$

13.3.1.2. Inference

The final goal of the PLSM model is to analyze temporal documents and automatically infer the motifs and when they appear. These elements to infer are the parameters of the model, denoted as Θ, and composed of $p(z|d)$, $p(t_s|z,d)$ and $p(w,t_r|z)$. The maximum likelihood estimator can be obtained by maximizing the log-likelihood of the observed data (denoted D). After marginalization of the hidden variables $Y = \{t_s,z\}$, the log-likelihood is written:

$$L(\Theta|D) = \sum_{d=1}^{D}\sum_{w=1}^{N_w}\sum_{t_a=1}^{T_d} n(w,t_a,d)\cdot \log \sum_{z=1}^{N_z}\sum_{t_s=1}^{T_{ds}} p(w,t_a,d,z,t_S,t_r) \quad [13.6]$$

An expectation-maximization (EM) algorithm can be derived from the expression of the log-likelihood and enables us to obtain the parameters of the model. The details of the EM procedure are available in [VAR 10]. In order to improve the quality of the results obtained, it is possible to integrate a sparsity constraint directly into the EM algorithm, i.e. a constraint aimed at minimizing the number of non-null values in the tables to be estimated.

13.3.2. *Motifs extracted by PLSM*

In this section, we illustrate the results obtained by PLSM applied to videos.

13.3.2.1. Creation of temporal documents

For reasons of computation costs, PLSM is not directly applied to the documents presented in the previous section. In order to simplify the observations, low-level documents over short time periods (typically 1 second) are first created and processed with PLSA using a high number of topics (typically 75) following the approach described in previous section. The PLSA probability of each topic at a given time (multiplied by the number of words in the low-level document) is used as an observation for PLSM. In this way, the size of the vocabulary is reduced from

several thousand to only 75 words. Similarly, we reduce the temporal resolution: a one-second window is used to reduce the frequency to one observation per second. Finally, for an hour-long video, the size of the table $n(w,t_a,d)$ would be 75×3600 (75 values for w, 3600 time-steps in a single document).

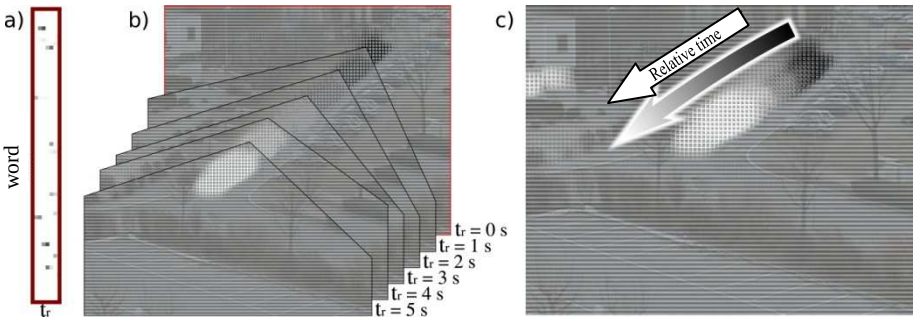

Figure 13.6. *Different representations for the motifs: a) in the form of a table; b) re-projecting each relative moment; c) using grayscale shading to represent time*

13.3.2.2. Representation of the motifs

A motif is a table which gives a probability for each word in the vocabulary at each relative time. The words in the vocabulary correspond to PLSA topics and therefore to regions of activity in the image. At each relative time, it is therefore possible to re-project the various words from the vocabulary into the image. An animation successively showing the relative moments can be used to view the motif over time. Using grayscale shading, it is possible to condense that animation into a single image to represent it on paper. Figure 13.6 illustrates these different representations.

13.3.2.3. Examples of motifs extracted by PLSM

Figures 13.7 and 13.8 show the representative motifs obtained on two different scenes. Generally, PLSM is able to extract the main activities in scenes which are presented to it.

Figure 13.7. *Dataset of a crossroads with pedestrians and cars. Top row: vehicle activity; bottom row: pedestrian activity*

Figure 13.8. *Dataset of a complex crossroads with cars and trams. The various activities of cars and trams are extracted correctly*

13.4. Applications: counting, anomaly detection

PLSM offers rapid comprehension of the scene thanks to the motifs that it extracts. Beyond this comprehension, it is possible to use the motifs and the times at which they appear in different ways. In this section, we show how the model can be used to perform counting on the one hand and anomaly detection on the other. Note that once the motifs have been learnt by the PLSM system, it is possible to save these motifs and find the times at which they appear in a new video.

13.4.1. Counting

The motifs extracted by PLSM correspond to recurrent activities in the scene. When a motif z represents a given event, it is possible to use the moments at which that motif appears to construct a detector for the corresponding event. We need only apply thresholding to $p(t_s|z,d)$ to create an event detector. We have created detectors in this way and evaluated their efficacy on videos for which we had annotated the ground truth. Figure 13.9 gives the precision/recall curves for three types of events. The curves are obtained by adjusting the value of the threshold applied to $p(t_s|z,d)$.

The results are excellent, and PLSM is able to detect the frequent events which it has learnt. Similar results have been obtained in the context of audio event detection.

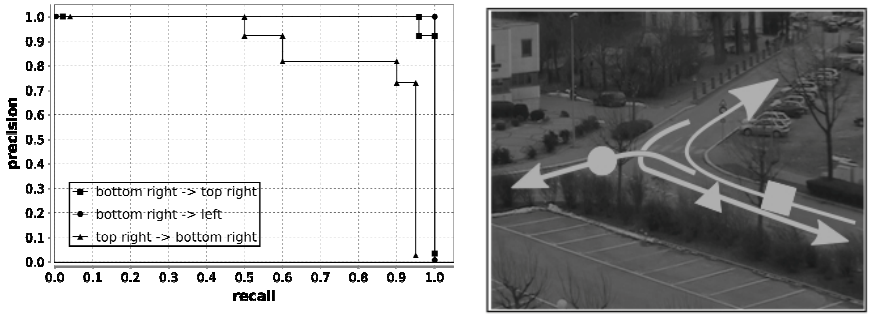

Figure 13.9. *Precision/recall curve for three types of events associated with three motifs over twelve minutes of video*

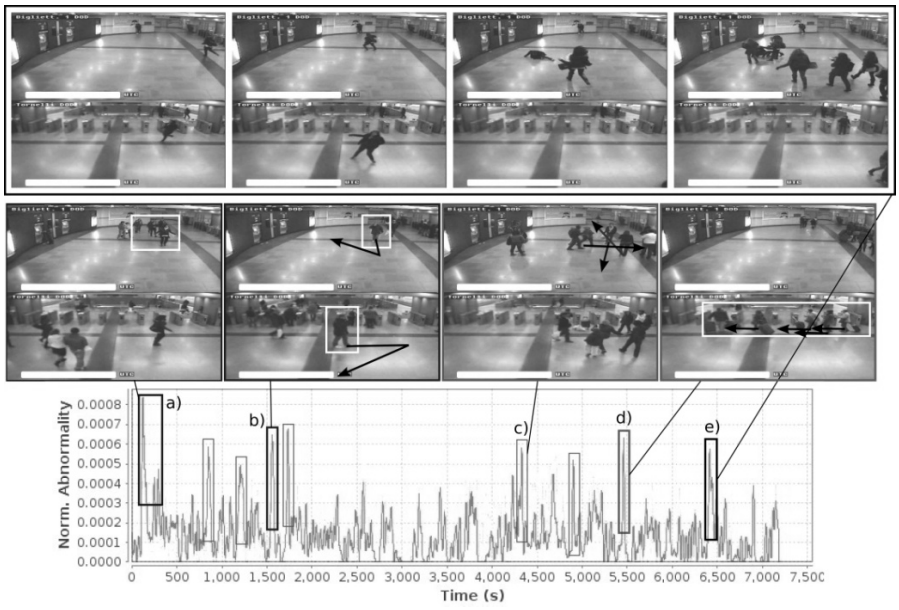

Figure 13.10. *Abnormality detection in a metro station*

13.4.2. Anomaly detection

PLSM captures recurrent activities, so the motifs represent normal (typical) things. If the motifs have already been learnt, it is possible to examine the extent to which these motifs are able to explain a new video of that scene. In order to do this, PLSM is used to find the moments at which fixed/saved motifs appear in the new video. Other measures can be conceived of (e.g. likelihood) but here we use reconstruction error as a measure of abnormality to quantify the extent to which the new video can be explained using these motifs. This error is defined thus:

$$\begin{aligned} &\textit{anormalité}(t_a, d) = \sum_w \left| \frac{n(w, t_a, d)}{n(d)} - p(w, t_a | d) \right| \\ &\text{avec}: p(w, t_a | d) = \sum_{t_s} \sum_z p(t_s, z | d) \cdot p(w, t_r = t_a - t_s | z) \end{aligned} \quad [13.7]$$

The reconstruction error is calculated at each time t_a. We can apply a threshold to construct an anomaly detector. Figure 13.10 shows an abnormality curve produced by PLSM applied to a pair of cameras in a metro station. The abnormality peaks above a certain threshold are illustrated by shots extracted at the corresponding times. Some of the abnormalities detected are due to groups moving in an unusual way (Figures 13.10a and 13.10d). Most are caused by atypical trajectories of people due to congestion in the station (Figure 13.10c and all the other rectangles illustrated by the fine line on the graph). A significant anomaly is detected (Figure 13.10e): a person runs in, following a circular path; then falls and is joined by a group of people who come to her aid.

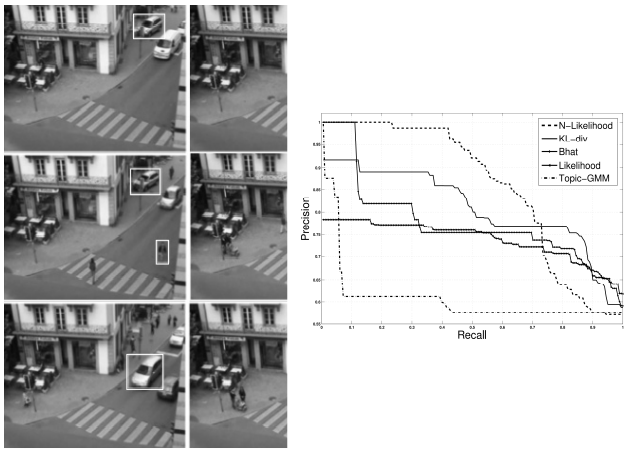

Figure 13.11. *Abnormality from PLSA: six examples of abnormality in the scene under consideration, and precision/recall graphs obtained from different abnormality measurements (derived from the PLSA model)*

It is interesting to note that PLSA (without a time dimension) is already able to detect numerous abnormalities when this detection does not require temporal reasoning. Such anomaly detection using PLSA is studied in detail in [VAR 09], where different measures of abnormality are compared as illustrated in Figure 13.11.

13.4.3. Sensor selection

Beyond pure anomaly detection, abnormality measurement can be used to preselect the sensors (cameras and microphones) to be displayed to an operator for interpretation. Indeed, it is impossible to constantly monitor the majority of the cameras installed in a metro network, for lack of staff; pre-selection of the cameras showing a high degree of abnormality is therefore highly advantageous to improve security.

13.4.4. Prediction and statistics

Models which incorporate time-related information such as PLSM can be used for short-term prediction of observed activities: when an activity is begun, it is possible to assume how it will end. A statistical analysis of the occurrences of activities captured by PLSM also enables us to make longer-term predictions. For instance, the usage of a metro station can be studied and predicted; here, topic models serve to extract high-level descriptors (e.g. the occurrences of a typical activity).

13.5. Conclusion

This chapter has presented approaches based on “bags of words” and “topic models”. Generally speaking, these approaches are particularly well adapted and effective to carry out unsupervised extraction of the main activities contained in a scene.

By selecting a vocabulary (for the words, position and orientation of movement, for instance) and a topic model such as PLSA or PLSM, it is possible to capture different kinds of information in the topics. In particular, the PLSM model can be used both to extract topics/motifs containing a large amount of temporal information, and to determine when these motifs appear.

Topics, extracted in a completely unsupervised manner by the approaches presented herein, offer us a very concise summary of the activities present in a

scene. Beyond comprehension of scenes, this chapter has also presented how a model such as PLSM can be successfully used to count frequent events or detect abnormal activities.

Because of the quality of the results obtained and the wide domains to which they are applicable, topic models have a definite future in the field of activity recognition in videos and multimedia documents.

Acknowledgments

The authors gratefully acknowledge the financial support from the Swiss National Science Foundation (Project: FNS-198,HAI) www.snf.ch/E and from the 7th framework program of the European Union project VANAHEIM (248907) www.vanaheim-project.eu under which this work was done.

13.6. Bibliography

[BLE 03] BLEI D.M., NG A.Y., JORDAN M.I., "Latent Dirichlet allocation", *Machine Learning Research*, n° 3, p. 993-1022, 2003.

[FAR 08] FARRAHI K., GATICA-PEREZ D., "What did you do today? Discovering daily routines from large-scale mobile data", *ACM Multimedia*, Vancouver, Canada, 2008.

[HUY 08] HUYNH T., FRITZ M., SCHIELE B., "Discovery of activity patterns using topic models", *Ubiquitous computing (UbiComp)*, p. 10-19, Seoul, South Korea, 2008.

[HOF 01] HOFMANN T., "Unsupervised learning by probability latent semantic analysis", *Machine Learning*, n° 42, p. 177-196, 2001.

[NIE 08] NIEBLES J.C., WANG H., LI F.F., "Unsupervised learning of human action categories using spatial-temporal words", *IJCV*, vol. 79, n° 3, p. 299-318, 2008.

[VAR 09] VARADARAJAN J., ODOBEZ J.M., "Topic models for scene analysis and abnormality detection", *ICCV-12th International Workshop on Visual Surveillance (VS)*, Kyoto, Japan, 2009.

[VAR 10] VARADARAJAN J., EMONET R., ODOBEZ J.M., "Probabilistic latent sequential motifs: Discovering temporal activity patterns in video scenes", *BMVC*, Aberystwyth, United Kingdom, 2010.

[WAN 09] WANG X., MA X., GRIMSON E.L., "Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models", *IEEE Transactions on PAMI*, vol. 31, n° 3, p. 539-555, 2009.