Detecting queues at vending machines: a statistical layered approach

Xavier Naturel, Jean-Marc Odobez*

IDIAP Research Institute, P.O. Box 592, Centre du Parc, 1920 Martigny, Switzerland
{*xnaturel,odobez*}*@idiap.ch*

Abstract

This paper presents a method for monitoring activities at a ticket vending machine in a video-surveillance context. Rather than relying on the output of a tracking module, which is prone to errors, the events are directly recognized from image measurements. This especially does not require tracking. A statistical layered approach is proposed, where in the first layer, several sub-events are defined and detected using a discriminative approach. The second layer uses the result of the first and models the temporal relationships of the high-level event using a Hidden Markov Model (HMM). Results are assessed on 3h30 hours of real video footage coming from Turin metro station.

1 Introduction

Our overall objective is to monitor the general usage of a metro station equipment, with an emphasis on ticket vending machine (machine usage, machine mis-use or vandalism, anomalies), and extract their statistics. It is expected that automatic generation of statistics of the station's usage will provide a better understanding for metro operators. We focus on the recognition of one specific event (queues), to illustrate our approach, but some other events (e.g. machine occupancy) are also detected, which can provide interesting insights.

Most of the previous works on human activity recognition are based on tracking [1], where the tracks are further analysed using HMM [5], bayesian networks [4] or clustering [8]. However, tracking is computation intensive and can perform badly with even medium crowding situation. Another approach is to rely instead on intensity measure and/or optical flow [2, 3]. To detect complex events in video-surveillance, ontologies are often used [7, 4]. Ontologies are useful to build scenarios, that are usually recognized through a set of

rules. While generic, these approaches rely on the exact extraction of the entities, and thus usually break down when a failure occur in this process (e.g. tracking failure).

One contribution of this work is to model events directly from image measurements. To this end, it is proposed to use a statistical layered model using features derived from background subtraction. Previous approaches in computer vision using a layered model have been used to recognize office activities [6] or group activities in meetings[10]. The layered approach is known to require less training data, with equal performance [6]. Another advantage is that the second layer is quite robust to changes in the raw features, since it only depends from abstracts events, which are supposedly more stable than the raw features. Previous approaches were using HMM for both layers. As a second contribution, we propose to use Support Vector Machine (SVM) for the first layer, to be able to use high-dimensionnal features, and to reduce the amount of training data.

In the next sections we will defined the task more precisely, define the features and the recognition model for the sub-events in section 3, and the modelling of the queuing event in section 4. Section 5 presents the results for sub-event recognition, queue recognition, as well as some statistics for machine occupancy.

2 Task definition and overall approach

A queuing event is defined as people waiting to access the vending machine. It is defined by temporal relationships between sub-events: some people are waiting while the machine is occupied; then people who were using the machine are leaving, and people are entering the machine zone to use it, while there still are people waiting. As a consequence, queuing can be defined by a set of 4 sub-events. These are: { *machine occupancy (people in zone* $\mathbf{z}_m$*), entering zone* $\mathbf{z}_m \cup \mathbf{z}_w$*, leaving zone* $\mathbf{z}_m \cup \mathbf{z}_w$*, people waiting in zone* $\mathbf{z}_w$ }, where $\mathbf{z}_m$ denotes a region just in front of the vending machine, while $\mathbf{z}_w$ corresponds to a farther region in the hall (see figure 2).

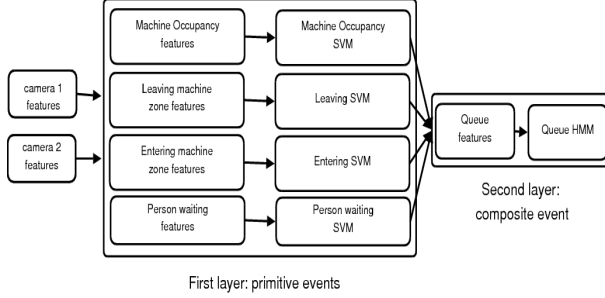To recognize a queuing event, we propose to use a

---

Figure 1. The layered approach.

layered approach, as described in figure 1. In the first layer, separate modules detect the primitive events, and the classification results of these primitive events are fed to the second layer. In [6], the first layer is built using HMM models. However, in our case, the feature set is quite large and we do not have a large number of positive example for training data. Using a generative model like HMM to do recognition might not be appropriate. We thus propose to rely on a discriminative approach instead, specifically SVM.

3 Primitive events recognition

3.1 Feature definition

To compute the features, a set of 3D cylinders is defined on the ground plane, roughly corresponding to an average human height and width (1m80 high and 80 cm wide). These 3D cylinders are then projected into the image plane, using calibration. This results in a set of 2D bounding boxes. The set of boxes is composed of

Figure 2. Bounding boxes set-up.

a first row of 3 boxes (zone $\mathbf{z}_m$), and a grid of 3 by 5 boxes behind this front row, in the waiting zone $\mathbf{z}_w$, as illustrated in figure 2, which shows the boxes setup in one camera view, as well as a schematic view of the boxes position, and the machine and waiting zones $\mathbf{z}_m$ and $\mathbf{z}_w$.

Background subtraction using the technique from [9] is performed on each camera view, and the resulting images are binarized. For each box, the percentage of foreground pixels in this particular box is computed, which

can be interpreted as the correlation between the foreground image and the binary mask template defined by the box. The correlation in box $1 \leq i \leq 18$, in camera $1 \leq j \leq 2$, at time instant t will be denoted by $b_i^j(t)$. This set of features are the core features from which all the features used to recognize the primitive events will be defined, as described below.

Machine occupancy features: Only 3 boxes next to the machine are considered. The feature vector F_{mo} is defined as:

$$F_{mo}(t) = \{b_i^j(t), b_i^j(t+1), b_i^j(t+2)\}, i = 1, 2, 3, j = 1, 2$$

Leaving/entering features: All the boxes are used ($i = 1 \ldots 18$), and the feature vector has 108 dimensions:

$$F_{le}(t) = \{b_i^j(t), bv_i^j(t+1) - bv_i^j(t), bh_i^j(t+1) - bh_i^j(t)\}$$

and $bv_i^j(t) = b_i^j(t) - b_{v(i)}^j(t), \quad bh_i^j(t) = b_i^j(t) - b_{h(i)}^j(t)$

with $h(i)$ and $v(i)$ are respectively one of the nearest box in horizontal and vertical directions on the ground plane for box i (e.g. in figure 2, $v(1) = 2$, $h(1) = 6$, and $v(5) = 4$, $h(5) = 10$).

Waiting people features: In that case, we have one feature vector per box, as the goal is to detect if a person is waiting near a psotition i:

$$F_w^i(t) = \{m_i^j(t), m_i^j(t + 3), m_i^j(t + 6), \quad (1)$$

$$|m_i^j(t) - m_i^j(t + 3)|, |m_i^j(t + 3) - m_i^j(t + 6)|\} \quad (2)$$

$$\text{with } m_i^j(t) = \frac{1}{3} \sum_{k=t}^{t+2} b_i^j(k) \quad (3)$$

These feature vectors try to capture the spatial and temporal characteristics of each sub-event with appropriate temporal and spatial derivatives, as well as using several time instants, so as to represent motion, and increase robustness with respect to ambiguous cases, like someone walking in front of the machine.

3.2 Recognition models

A classification algorithm is applied on each feature vector F_{mo}, F_{le}, F_w, to classify each frame. SVM have been chosen because they are known to handle well large dimensional input spaces, and potentially small amount of training data, as it is the case here. In this work, we use a SVM with soft-margin and a Gaussian kernel, $K(\mathbf{b}_1, \mathbf{b}_2) = \exp^{-\frac{||\mathbf{b}_1 - \mathbf{b}_2||^2}{2\sigma^2}}$. SVM parameters (bandwidth σ, and regularization cost C) are obtained through cross-validation, by an exhaustive search of the parameter space. To produce the hard output, the soft SVM output is thresholded using the value that maximizes the F-measure on the training set.
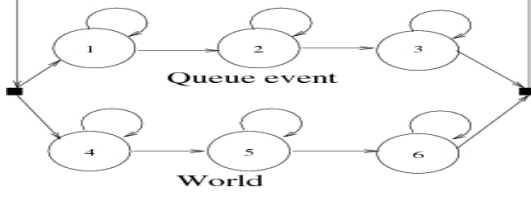
Figure 3. HMM state model for queue detection

4 Modelling queuing events

4.1 Baseline

We first propose a straightforward approach to model queuing events, where raw features are directly used. The model is shown on figure 3. States 1, 2 and 3 form a left-right model which is the sequence of states we have to go through to recognize a queue event. States 4, 5 and 6 act as a world model, describing other data sequences. Observations are modeled by a Gaussian Mixture Model (GMM) with 4 mixtures. As features, we use those defined for the leaving/entering event, $F = F_{le}$, as they are defined on the whole space $\mathbf{z}_m \cup \mathbf{z}_w$, and include temporal and spatio-temporal information that are likely to be useful to detect our pattern of interest.

4.2 Layered Model

Features: The principle is to use the SVM output of the first layer as input to second layer. In order to have homogeneous values between the SVM scores, a normalization procedure is applied on each raw SVM output y_i of a given sub-event i. A median filter of size 3 is first applied to temporally smooth the signal, and the sigmoid function $g(y_i) = \frac{1}{1+e^{-\lambda(y_i-t_i)}}$ is then applied to obtained normalized values, where t_i is the threshold that maximizes the F-measure on the training set of each sub-event i. The input feature vector is then defined as $g = \{g(y_{MO}), g(y_L), g(y_E), g(y_W)\}$. For *Waiting*, the classifier is applied on each box of the zone $\mathbf{z}_w$, we thus have as many outputs as number of boxes in the zone $\mathbf{z}_w$. To solve this, y_W is simply chosen as the maximum value over all boxes, at each time instant. An example of four normalized features is plotted on figure 4, in which someone is going to the machine, buying a ticket, and then leaving.

Recognition model: The normalized feature vector $g(y)$ is used as input to an HMM. The architecture is the same as in the baseline model (fig. 3), only the inputs are different. Observations are also modeled by a GMM with 4 mixtures. Note that this is the same model architecture as the baseline, .
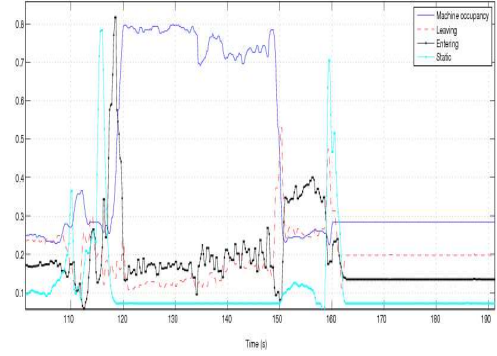

Figure 4. Outputs of the SVM classifers after normalization.

5 Results
5.1 Datasets and performance measures

The evaluation has been conducted on two videos from Turin metro station, of a total duration of 3h30. For all primitive events as well as for queues, the number of frames in which events are present (positive frames) is very small (1 to 5% of the total stream). Instead of computing a global classification rate, measures are computed **only on the positive events**, e.g. we count all the frames correctly detected as a positive event N_g (good detection), all the negatives frames detected as positive N_f (false positive), and all the positive frames detected as negative N_m (missed). Precision, recall, and F-measure are given by:

$$P = \frac{N_g}{N_g + N_f} \quad R = \frac{N_g}{N_g + N_m} \quad F = \frac{2PR}{P + R}$$

. We also define a similar measure for the event, where an event is a set of contiguous positive frames detections. Detected events are matched against the ground truth using a dynamic time warping procedure with temporal overlapping constraints.

5.2 Sub-events results

The results of Occupancy, Leaving and Entering events are presented in table 1. While detecting machine occupancy is not so difficult, we still have to cope with people walking in front of the machine, and not using it. Results are also dependant on the quality of the background subtraction. Note that the recognition of this event using an HMM approach results in a F-measure of 91 (frame) and 92.7 (event), justifying our choice of a SVM instead of a HMM, as usually done in the multilayer approaches of [6, 10].

In the case of Leaving/Entering, results in terms of events are quite good, most leaving and entering events are detected. False alarms and missed detections

	Frame			Event		
	P	R	F	P	R	F
Occupancy	91.8	99.8	95.6	94	96.9	95.4
Entering	74.2	57.3	64.7	87	90.9	88.9
Leaving	66	49.1	56.3	93.7	78.9	85.7

Table 1. Sub-events results

are mainly due to the difficulty of defining the events boundaries, what a leaving or entering event is, i.e. people wandering around the zone, or entering/leaving it slowly step by step.

In table 2, we also present some statistics of machine occupancy detection. Results seem to fit the ground truth quite well.

Statistic	Ground truth	Estimated value
Number of events	27	27
Mean duration(s)	32.5	34.9
Max duration (s)	1	6.8
Min duration (s)	99	94.6

Table 2. Statistics of machine occupancy versus ground truth

5.3 Queue Detection

	Frame			Event		
	P	R	F	P	R	F
Baseline	11.5	100	20.6	11.7	87.5	20.6
Layered	54.3	99.9	70.4	33.3	100	50

Table 3. Queue detection results

An example of queuing event is shown on figure label 5. Results of both baseline and layered approach are shown in table 3. The baseline results are very low, with a very high number of false alarms. The model is not able to distinguish between a queue event happening, and the mere presence of people in the zone. Note that the frame-based recall is 100% but the event-based recall is lower. This can be explained by the fact that several queueing events are detected as a single event only, thus only one event is counted as correctly retrieved. The model might be disturbed by a lot of people moving around the zone, while not actually queuing. In the layered approach, results are better than in the baseline, although quite a high number of false alarms are still present. Most false alarms are occuring when several people are present in the waiting zone $\mathbf{z}_w$.

6 Conclusion and perspectives

We have presented an event recognition based approach for monitoring ticket vending machines. A

Figure 5. Example of a queuing event

model for detecting high level events with a layered architecture has been presented. The high level event is composed of several sub-events, which are detected by SVM. These sub-events are also of interest and can provide useful information to monitor the station's activity. Classification results for the queuing event show that the layered approach outperforms a baseline one with raw features, but still need improvement. A more constrained architecture of the HMM could help to discriminate the queue events.

References

[1] J. K. Aggarwal and Q. Cai. Human motion analysis: a review. *Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE*, pages 90–102, 1997.

[2] E. L. Andrade, S. Blunsden, and R. B. Fisher. Modelling crowd scenes for event detection. In *ICPR'06*, volume 1, pages 175–178, 2006.

[3] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV'03*, page 726, Washington, DC, USA, 2003.

[4] S. Hongeng, R. Nevatia, and F. Bremond. Video-based event recognition: activity representation and probabilistic recognition methods. *Comput. Vis. Image Underst.*, 96(2):129–162, 2004.

[5] V. Nair and J. Clark. Automated visual surveillance using hidden markov models. In *VI02*, 2002.

[6] N. Oliver, A. Garg, and E. Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. *Comput. Vis. Image Underst.*, 96(2):163–180, 2004.

[7] L. Snidaro, M. Belluz, and G. Foresti. Representing and recognizing complex events in surveillance applications. In *AVSS'07*, pages 493–498, London, UK, 2007.

[8] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE PAMI*, 22(8):747–757, 2000.

[9] J. Yao and J.-M. Odobez. Multi-layer background subtraction based on color and texture. In *CVPR-VS workshop*, Minneapolis, June 2007.

[10] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud. Modeling Individual and Group Actions in Meetings: a Two-Layer HMM Framework. *IEEE Transactions on Multimedia*, 8(3):509–520, 2006.