Journal of Information Assurance and Security.
ISSN 1554-1010 Volume 11 (2016) pp. 331-339

M-SVM classifiers for Events detection in Video using Auto-associative Neural Network Models

Mohamed Chakroun, Yassine Aribi, Ali Wali and Adel M. Alimi

REGIM: Research Groups on Intelligent Machines, University of Sfax,
National Engineering School of Sfax (ENIS),
BP 1173, Sfax, 3038, Tunisia
mohamed.chakroun.tn@ieee.org, yassine.aribi.tn@ieee.org, ali.wali@ieee.org, adel.alimi@ieee.org

Abstract: **Our main goal in this study was to develop and validate an intelligent system for video event detection based on spatiotemporel features combining an auto-associative neural network models for feature reduction. Proposed system aims at high accuracy of event classification mainly with the use of an M-SVM model. The core of the system is the auto-associative neural network models which can reduce the size of feature vectors. The proposed model performance, evaluated an important data basis including seven events, was compared to other models found in the literature; it outperforms the other methods in terms of precision.**

Keywords: Event detection, M-SVM, HOG/HOF features, auto-associative memory,neural network, modeling.

I. Introduction

Event detection applications, characterized by their complexity and unexpected aspect, they require the development of a number of specic functionality and modular components that are specialized to solve this particular problem.
The event detection can be used by a video surveillance system that allows the movements to be tracked automatically based on the change between consecutive images of a video stream and the properties of the object such as: size, position and speed.
In another hand it is impossible to address the automatic indexing in the same way as the manual one. Indeed the manual indexing is done by a librarian who can determine all the design elements that determine the purpose and the theme of a document. A computer, however, can not carry out this conceptual analysis automatically; and even if it tries, it would not be with the same depth. Automatic indexing is ultimately based on the signal processing associated with meaningless digital information. In most of the cases this signal does not contain any contextual information.
Image and video indexing and retrieval continue to be an extremely active study area within the broader multimedia research community [17].
The main goal of the current study is the implementation of a new system for event detection in video based on spatio-temporal feature such as HOG and HOF descriptors using an auto-associative memory.
This paper is organized as follow: In section 2, we present the related works to this field. Section 3 we introduce an overview of our adopted approach with the detailed used techniques and modules. In section 4 we evaluate the performance of our approach with comparison according to other existent techniques. Finally, section 5 concludes our presented work.

II. Related works

The multiplication of the video data available and the generalization of surveillance cameras and have generated the need for generics video indexation tools capable of performing the semi-automatic search events in a video database. In this context, a first intelligentt approach is to be extracted from the video data of the attributes spatial and temporal sufficiently compacts and reliable to effectively represent the events that happen in the video sequence.
There are many approaches have been proposed in the literature of which we can cite in particular the spatial and temporal attributes colorimetric extraction combined with statistical modeling techniques[8, 4] , such as Hidden Markov Models (HMM) [3, 10].
The modeling by kernel is used also in [5] to construct a modeling of objects sought. A system of event detection for the videos surveillance has been developed [7] in the context of TRECVID 2009, which can able to detect the following three events: ElevatorNoEntry, OpposingFlow and PersonRuns.
Another system of event detection for video surveillance called “ESUR” [23] is well presented in the literature. The data of this system comes from five different requiring cameras , for each camera, there is a generation of a reference model. The human body is integrated for the final detection results in the step of detecting an object.
Different approaches of presenting the competition TRECVID 2010, in fact the system proposed in [15] for event detection consists of three steps . The detection and tracking points of interest is the first step, the feature extraction is the second step and the classification based on SVM is the third step.
A Bayesian network to extract the interesting moments in the Formula 1 videos was developed by the authors of [12] . Another system was developed using the Bayesian networks

for the detection of events in the sports videos [18] . In [2], the authors presented a new approach for event detection from video surveillance data based on optical flow histogram with no prior knowledge of the motion nature.
In [24], the authors presented a new method for event detection using a new approach for event detection from video surveillance system based on incremental learning.
In[1] the authors used the classic linear SVM classifiers to classify and identify different events. However, eSur trecvid 2010 used SVM-HMM [21].
In the literature, many event detection systems are based on classical learning. In this paper we propose new event detection based on an auto-associative neural network model for feature reduction phase and Multi-SVM (M-SVM) for classification phase. Few authors have used the concept of auto-associative neural network in conjuction of Multi-SVM that has given more satisfactory results [25]. In fact, in [13] different types of exciting events in a broadcast soccer video is detected using the M-SVM. In [16, 17] the authors have proposed a multi-SVM incremental learning system based on Learn++ classifier for the detection of predefined events in the video.
In [9] the authors have proposed an anomaly-detection approach applied for video surveillance in crowded scenes. This approach is an unsupervised statistical learning framework based on analysis of spatio-temporal video-volume configuration within video cubes.
In 2016 the authors of the paper [26] proposed a reliable visual analysis technique for fast fire flame detection in surveillance video using logistic regression and temporal smoothing. As the above mentioned challenges are met, and experience is gained, implementation of validated techniques in commercial software packages will be useful to attract the interest of the computer vision community and increase the popularity of these tools. The application of event detection in video is well established in research environments and is still limited in computer vision settings to institutions with extensive computing support. It is expected that with the availability of computing power in the near future, more complex and ambitious computer intensive event detection systems will become feasible.We are going to describe the algorithms and the techniques retained in our contributions later.

III. PROPOSED SYSTEM

In many machine learning approaches, the acquisition phase of a representative set of training data is often long and expensive. In our context, a new classifier must be created and the time required to build a representative data set must be minimized so that it is acceptable from the point of view of the user. For this, the classifier must be learned quickly with few examples, and then be incrementally updated by each new example available. Our system includes four phases, the video segmentation, the features extraction and reduction based on auto-associative neural network, learning phase based on M-SVM for event detection and classification. The system diagram is illustrated in Figure 1. Each phase will be described in details.

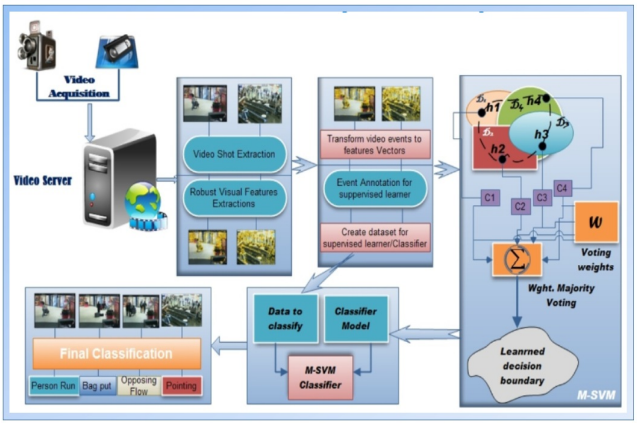

Figure. 1: Overview of our system

A. *Video segmentation phase*

Once we get the video to process, we will divide this sequence into a hierarchy of plans using an algorithm based mainly on the changes in movements. For each event, we manually segmented the input video into positive and negative sub-sequences. A positive example is like “event of interest is present one or more times in this video” and a negative example is like “event of interest is not present in this video”.

B. *Features extraction and reduction phase*

Afterwards, we apply the algorithms to generate the feature vectors for each image and prepare the learning phase. Then we segment the sub-sequences into image planes which contain a number of frames. In our case and for reasons of normalization, we randomly select 5 frames from each plane.
In Figure 2 we can see the diagram of the segmentation phase.

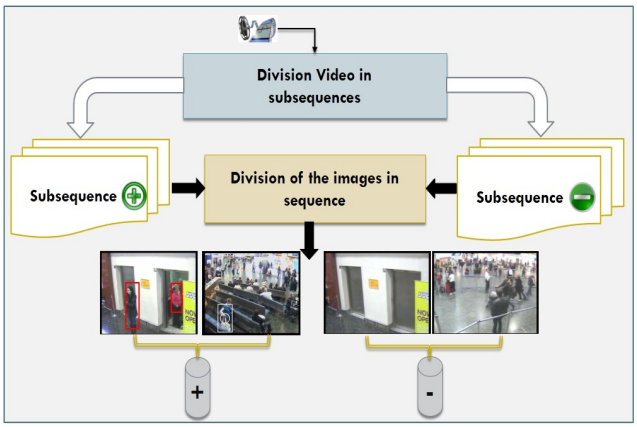

Figure. 2: Segmentation phase

- Positive Images: Showing one state in the set of states constituting the event.

- Negative Image: That contradicts the desired state.

1) Feature extraction

In our study, the feature vectors that correspond respectively of HOG (Histogram of Oriented Gradient) feature vector and HOF (Histogram of Optical Flow) feature vector are combined in a single vector using a simple concatenation that serves as input for the auto-associative memories network. The input layer contains (size of HOG vector+size of HOF vector)*number of selected frame=(72+90)*5 unit which equals the features vector dimension. Second layer contains in turn 1620 units, and the compression layer (layer 3) contains 70 units which was set empirically. Then the data in the layer 3 are used as input of an M-SVM classifier which provided the classification results.

HOG algorithm Scanning on the input image is based on detection window (64x128 pixels). The window is divided into cells (8x8 pixels), for each cell accumulating a histogram of gradient orientations over the pixels of the cell. For better invariance to illumination, histogram normalization can be done by accumulating a measure of the local histogram energy over blocks and using the results to normalize all cells in the block. The normalized histograms (HOG features) are collected over the detection window.

HOF algorithm HOF is built by considering the same neighborhood of point of interest obtained in the HOG algorithm by performing a histogram on the orientations of the optical flow extracted from consecutive selected frame. The histogram is obtained from 5 bins describing the movement to the right, left, up, down and no movement.

HOG/HOF algorithm The HOG/HOF descriptors were introduced by Laptev et al. in [19]. To characterize local motion and appearance, the authors compute histograms of spatial gradient and optic flow accumulated in space-time neighborhoods of detected interest points. For the combination of HOG/HOF descriptors with interest point detectors.Each volume is subdivided into a (n_x * n_y * n_t) grid of cells; for each cell, 4-bin histograms of gradient orientations (HOG) and 5-bin histograms of optic flow (HOF) are computed. Normalized histograms are concatenated into HOG, HOF as well as HOG/HOF descriptor vectors and are similar in spirit to the well known SIFT descriptor. In our evaluation we used the grid parameters n_x,n_y = 3, n_t = 2 as suggested by the authors.

Feature vector reduction In this work we use two parameter spaces (HOG and HOF features) to represent event, which gives a large feature vector that may affect the classication phase. To overcome this drawback we use a features reduction method based on auto-associative memories. It is considered for some work as a non-linear PCA (principal component analysis) [22, 6, 11] because they can retain only nonlinearly correlated primitives. The auto-associative memories are feedforward neural network used to capture the input data.
These networks consist of ve layers. The third layer contains the smaller number of units, it serves as a compression layer. More explain are given in figure 4.

Auto-associative memories serve to reproduce the input data (x1,x1,..,xn) in the output layer (x̂1,x̂2,.. x̂n) which is called identity mapping [20].
In fact, this process has two steps. First input data are reduced in the bottleneck layer it is an encoding phase. Second, data are decompressed from bottleneck layer to the output, it is a decoding phase. Learning process is stopped when the output data is closest as possible to the input. Encoding process is given by:

$$v_k = \sum_{j=1}^{N2} w_{2jk}\sigma\left(\sum_{i=1}^{N1} w_{1ij}x_i + \theta_i\right), k = 1, ..., M \quad (1)$$

Where, $N1$ is the number of input vectors, $N2$ is the number of mapping layers and M is the number of bottleneck layers. w is the weight value of the network and θ_i is the threshold value for the ith node of the mapping layer. σ is the sigmoid function and given as:

$$\sigma(x) = \frac{1}{1+e^x} \quad (2)$$

Decoding process is given by:

$$\hat{x}_k = \sum_{j=1}^{N2} w_{4jk}\sigma\left(\sum_{i=1}^{N1} w_{3ij}v_i + \theta_i\right), k = 1, ..., N1 \quad (3)$$

The training process is stopped when the error E is minimized. E is calculated as

$$E = \sum_{n=1}^{N}\sum_{i=1}^{N1}(x_{in} - \hat{x}_{in})^2 \quad (4)$$

where, N is the number of the training samples.

C. The learning phase using M-SVM

M-SVM is based on Learn++ algorithm. This latter, generates a number of weak classifiers from a data set with known label. Depending on the errors of the classifier generated low, the algorithm modifies the distribution of elements in the subset according to strengthen the presence of the most difficult to classify. This procedure is then repeated with a different set of data from the same dataset and new classifiers are generated. By combining their outputs according to the scheme of majority voting Littlestone we obtain the final classification rule.
The weak classifiers are classifiers that provide a rough estimate - about 50% or more correct classification - a rule of decision because they must be very quick to generate. A strong classifier from the majority of his time training to refine his decision criteria. Finding a weak classifier is not a trivial problem and the complexity of the task increases with the number of different classes, however, the use of NN algorithms can correctly resolved effectively circumvent the problem. The error is calculated by the equation:

$$error_t = \sum_{i:h_i(x_i)\neq y_i} S_t(i)\left[|h_t(x_i) \neq y_i|\right] \quad (5)$$

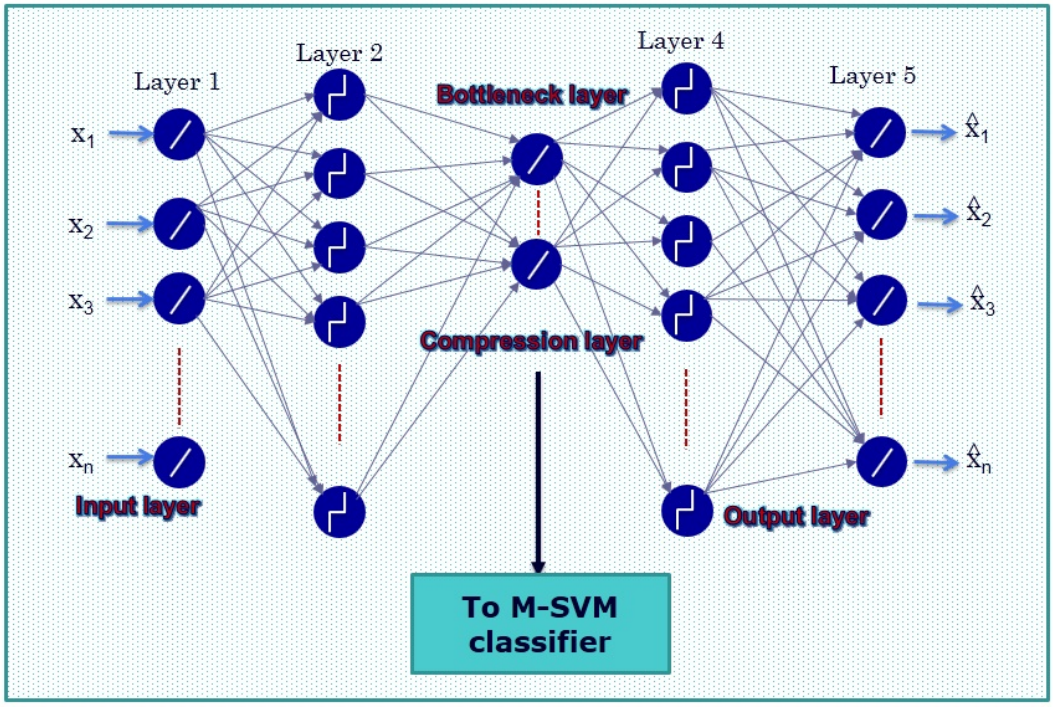

Figure. 3: Auto-associative Neural Network architecture

with $h_t : X \rightarrow Y$ an hypothesis and where TR_t is the subset of training subset and the TE_t is the test subset. The synaptic coefficients are updated using the following equation:

$$w_{t+1}(i) = w_t(i) * \left\{ \begin{array}{l} \beta_t \; if \; H_t(x_i) = y_i \\ 1 \; else \end{array} \right\} \qquad (6)$$

Where t is the iteration number, B_t composite error and standard composite hypothesis H_t.

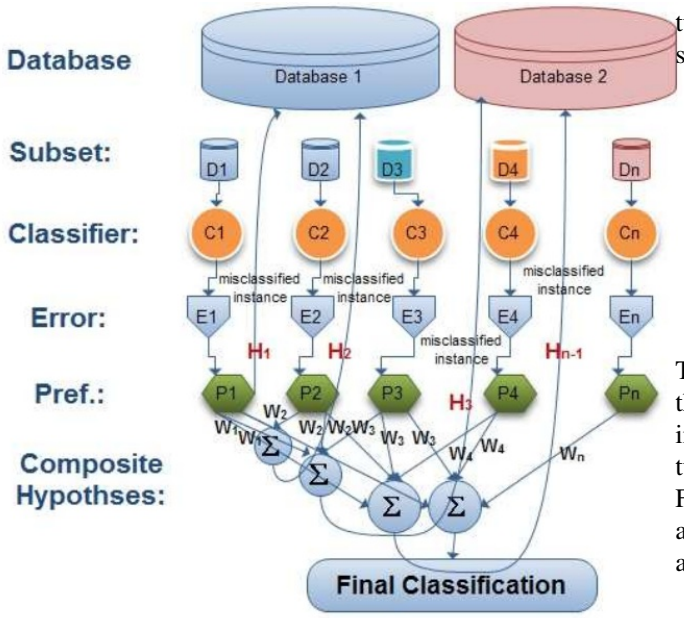

Figure. 4: M-SVM classifier

In our approach we replace each weak classifier by SVM. After T_k classifiers are generated for each D_k, the final ensemble of SVMs is obtained by the weighted majority of all composite SVMs:

$$H_{final}(x) = arg \underbrace{max}_{y \in Y} \sum_{k=1}^{K} \sum_{t:h_t(x)=y} log \frac{1}{\beta_t} \qquad (7)$$

D. Event detection phase

In our approach, an event is defined by a set of states. For detecting an event, firstly we should detect all states constituting this event. Our algorithm for detecting a state is described as follows: For each test image we will:

- Construct the feature vector of the image depending on the learning model.
- Verify the class membership of this image from the M-SVM.
- Decide if it is an image belonging to the set of states defining the event x_i.

The decision for belonging to a class is performed directly by the decision function of incremental SVM. Once we succeed in detecting all the possible states belonging to a set constituting an event, we can, at this moment, detect this event. Figure 5 illustrates the succession of states detected before arriving at the final state. The event will be detected if it has already detected all defined states constituting this event.

E. Event modelisation

We propose a description for each event, a definition of the starting time, a definition of the end time and modeling by finite state automaton. For the example of the event people SplitUp, the description of this event can be like as follows: Two or more persons, who position themselves, sit, move

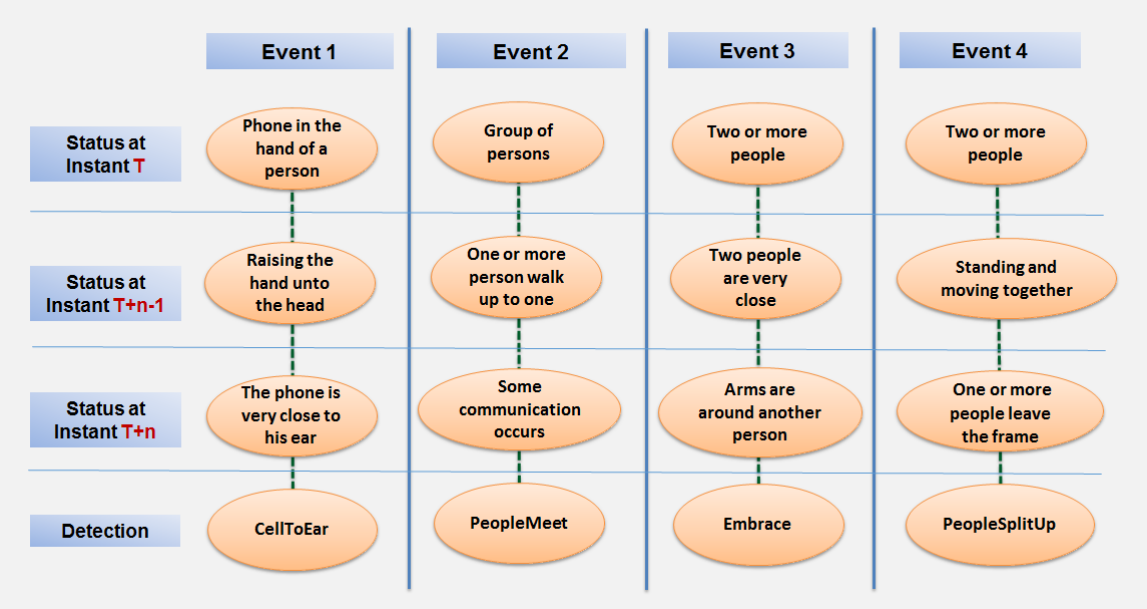

Figure. 5: Event detection phase

together or communicate. One or more persons who separate and leave the frame. Then the starting time is as follows: The last time that a group of persons is the closest one to each other. For the end time, it is the first time at least when a group member leaves the frame. In Figure 5 the finite state machine of the event People SplitUp is presented.

1) SVMQL "Surveillance Video Modeling Query Language"

In order to offer the user the ability to model his request for an advanced research, we developed a user interface with the ability to create a finite-state machine (FSM) made up of states (RVOs, RVEs or Frames) and transitions (which can be explicit transitions or events that lead to a stable state).

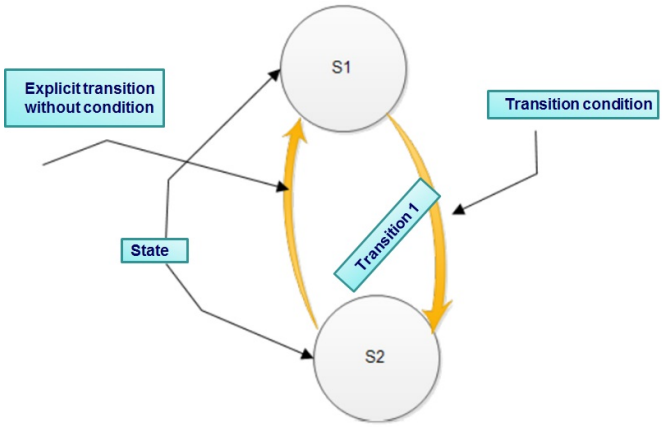

Figure. 6: Overview of an FSM

The user interface is the facet of an underlying proposed system that translates a visual query (FSM) to a language similar to SQL. This language is called SVMQL (Surveillance Video Modeling Query Language). Figure 7 and 8 shows two FSMs that the user want to nd respectively the event of Bag Put and the event of CellToEar.

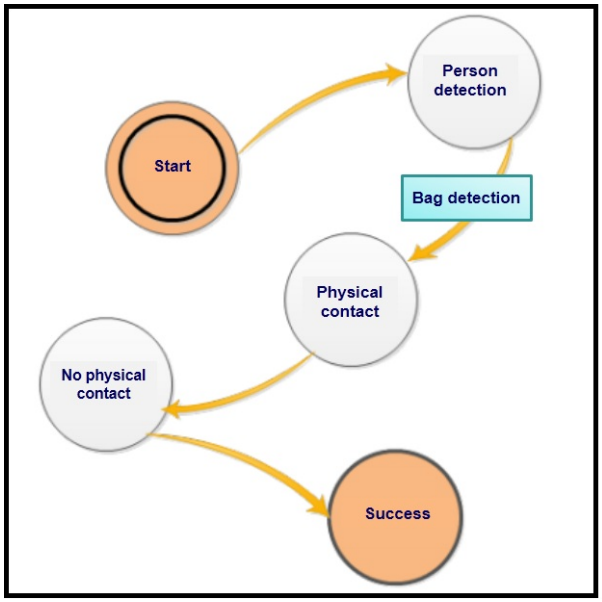

Figure. 7: Bag Put FSM

To design the visual query template the user selects the states and transitions from a list of predefined keywords.
Citing for example:

- person detection,
- group of people detection,
- displacement of a person,
- car detection,
- bag detection,
- open elevator,
- closed elevator,

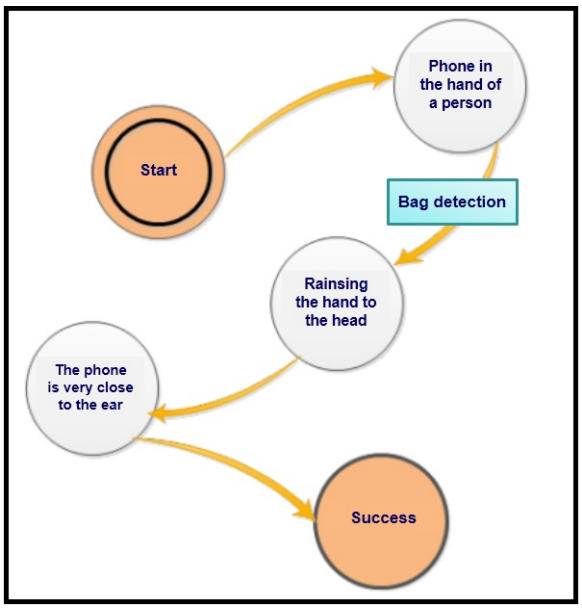

Figure. 8: CellToEar FSM

- physical contact,
- no physical contact,

2) Translation of FSM to simple query

For retrieve the surveillance video, we have proposed a system to translate a graphic query to query language. The syntax of a query expressed by this query language is the following: SELECT <Select item> FROM < Database-name> WHERE <Conditions>
Where: SELECT, FROM, WHERE are keywords for a query and they are mandatory :

- "Select item" species the returned results. It may be either whole Recognized Video Objects (or attributes) or whole Event (or attributes) or Frames. We have implemented a counter operator COUNT that counts the number of results returned.
- "Database-name" specifies which parts of the video database are used to check the <Condition>. It can be either $*$ for the whole video database or a list of named sub-parts. This is interesting for the surveillance video retrieval because the video database can be divided into several parts according to time or to location. It enables to accelerate the retrieval phase in the case the users know which parts of the video database they are interested in.
- "Conditions" specify the conditions the retrieved results must satisfy. The users express their requirements by dening this component.

For example, the "People SplitUp" FSM will be translated in the following query: SELECT e FROM $*$ WHERE ((e: RVO) AND (e.Name ="Group of Person")) AND (e gofrom "'Displacement of person"' to "'Disappearance of Persons"')
To answer this query, the system must perform the following steps:

1. At first it searches on full database the sub-sequences containing the items "group of person";
2. In the returned results, the system searches the sub-sequences containing a moving person ie the item "Displacement of Person";
3. Finally, the system looks, under sub-sequences found in the previous stage, the event "Disappearance of Persons".

IV. Experimental Results

Experiments are conducted on many sequences from TRECVid 2010 database of video surveillance and many other sequences from road trafic. About 30 hours are used to train the feature extraction system, that are segmented in shots. These shots were annotated with items in a list of 7 events. We use about 20 hours for the evaluation purpose. To evaluate the performance of our system we use, in a first time, the common measure from the information retrieval community : the Average Precision. Figure 10 show the evaluation of returned shots. The best results are obtained for all events.

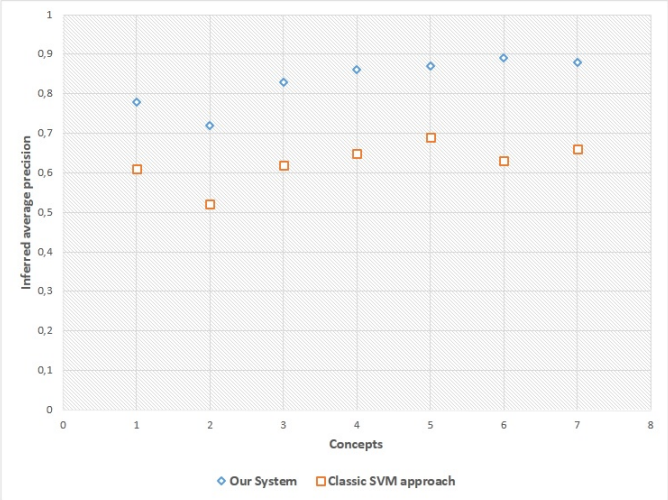

Figure. 9: Our M-SVM run score vs Classical SVM System

The dataset consists of surveillance camera which was acquired at London Gatwick airport; it is provided by TREC Video Retrieval Evaluation (TRECVid 2010) [14].
We seek to treat 7 events: "ObjectPut", "ElevatorNoEntry", "PeopleSplitUp", "PeopleMeet", "CellToEar", "PersonRun" and "Embrace" which are grouped into two categories: individual events and collaborative events.
The experiment is to apply our detection algorithm on a set of 2621 sub-sequences of the 7 events. The set of descriptors is used to supply the M-SVM classifier type that generates a model. For each event we have selected positive and negative examples to the learning phase.
In Figure 10 we note clearly the difference between the positive and negative examples in the event of "PersonRun".

In order to evaluate our event detection system, we have made a comparison between other systems, where #Ref is the number of annotated events, #Sys is the number of events

Figure. 10: Example of PersonRun event

detected, #Cordet is the number of correct detections, #Fa is the number of false detections and #Miss is the number of missed events.
Results in table 1 reveal that we have obtained a good performance in events PersonRun, CellToEar, Embrace, ElevatorNoEntry and PeopleSplitUp. However, the action ObjectPut, is relatively low.

V. Conclusion

In this paper we have proposed an event detection system in a video based on Multi-SVM classiers combining an auto-associative Neural Network Models. In our approach, an event is modeled by a set of chained statements constituting a controller, and in each state include a learning model with both positive and negative classes. The detection of an event will be ensured after the detection of all the states defining event. Our system includes four phases, the video segmentation, the features extraction and reduction based on auto-associative neural network, learning phase based on M-SVM for event detection and classification. The originality of our approach is the event detection manner using the minimum of features due to a reduction provided by the auto-associative neural network models. In the last part of this paper, we presented a solution for indexing video sequences by concept and predened event. In order to allow its use in many different contexts, the generic nature of the proposed solution is noteworthy. Our contribution stems from a set of constraints to respect and a set of problems to solve. In order to evaluate our approach, we apply it throughout a hundred hours of video recording in an airport and we have dealt the following seven events: “Embrace”,“ObjectPut”,“ElevatorNoEntry”, “PeopleSplitUp”, “PersonRun”, “CellToEar” and “PeopleMeet”. Our future work will focus on applying a cloud computing that can be applied in the computer vision eld to make it useful for the specific communities and may be extend this idea to android based mobile devices which might enable users to get more informations results via mobile devices quickly.

Acknowledgment

The authors would acknowledge the financial support of this work by grants from General Direction of Scientific Research (DRGST), Tunisia, under the ARUB program.

References

[1] A. Walha, A. Wali and A. M.Alimi. Developing a robust approach for detecting events in video streams, *Advanced Machine Learning Technologies and Applications Communications in Computer and Information Science*, 322, pp. 143-151, 2012.

[2] Wali and A. M.Alimi. Event detection from video surveillance data based on optical flow histogram and high-level feature extraction*In 20th International Workshop on Database and Expert Systems Application*, pp. 221-225, 2009.

[3] A. Bobick, A. Pentland and T. Poggio. Learning and Understanding Action in Video Imagery. In *Proceedings DARPA Image Understanding Workshop*, 1998.

[4] M. Chen, S. Chen and M. Shyu. Hierarchical Temporal Association Mining for Video Event Detection in Video Databases. *International Conference on Data Engineering Workshop*, 2007.

[5] P. H.Gosselin , M. Cord and S.Philipp-Foliguet. Kernel on Bags for Multi-Object Database Retrieval. *ACM International Conference on Image Processing, Atlanta USA*, 2006.

[6] J. Reyes, M. Vellasco and M. Tanscheit. Fault detection and measurements correction for multiple sensors using a modified autoassociative neural network. *Neural Computing and Application*, 2013.

[7] Y. Kentaro, T. Watanabe and S. Ito. Toshiba at TRECVID 2009: Surveillance Event Detection Task. *Corporate Research and Development Center, TOSHIBA Corporation*, 2009.

[8] G. Lavee, L. Khan and B. Thuraisingham . A Framework for a Video Analysis Tool for Suscpicious Event Detection. *Multimedia Tools and Applications*, 35:109-123,2007.

[9] Nannan Li, Xinyu Wu, Dan Xu, Huiwen Guo and Wei Feng. Spatio-temporal context analysis within video volumes for anomalous-event detection and localization. *Neurocomputing*, 155:309 - 319,2015.

[10] G. Papadopoulos, V. Mezaris, I. Kompatsiaris and M. Strintzis. Estimation and Representation of Accumulated Motion Characteristics for Semantic Event Detection. *International Conference on Image Processing (ICIP), San Diego, California, USA*,2008.

[11] E. Parviainen and D. Bottleneck. Classifiers in Supervised Dimension Reduction. *Artificial Neural Networks*,6354:pp.1-10,2010.

[12] M. Petkovic, V. Mihajlovic, W. Jonker and S. Djordjevic-Kajan. Multi-modal extraction of highlights from TV formula 1 programs. *Proc. of IEEE ICME*,pp.817-820,2002.

Table 1: Event Detection Scoring Analysis Report

	#Ref	# Sys	#Cordet	#Fa	#Miss
Our system	200	130	100	30	20
eSur @ trecvid 2010 SVM-HMM [15]	200	70	30	40	90
Classical SVM system [1]	200	150	50	100	70
Embrase Event					
	#Ref	**# Sys**	**#Cordet**	**#Fa**	**#Miss**
Our system	200	100	70	30	10
eSur @ trecvid 2010 SVM-HMM[15]	200	80	40	40	40
Classical SVM system [1]	200	120	50	70	30
ElevatorNoEntry Event					
	#Ref	**#Sys**	**#Cordet**	**#Fa**	**#Miss**
Our system	120	60	50	10	0
eSur @ trecvid 2010 SVM-HMM[15]	120	50	20	30	20
Classical SVM system [1]	120	100	28	72	12
People SplitUP Event					
	#Ref	**#Sys**	**#Cordet**	**#Fa**	**#Miss**
Our system	180	120	90	30	0
eSur @ trecvid 2010 SVM-HMM[15]	180	140	85	55	15
Classical SVM system [1]	180	134	60	74	40
PeopleMeet Event					
#Sys	**#Cordet**	**#Fa**	**#Miss**		
Our system	180	120	90	30	0
eSur @ trecvid 2010 SVM-HMM[15]	180	140	85	55	15
Classical SVM system [1]	180	134	60	74	40
CellToEar Event					
	#Sys	**#Cordet**	**#Fa**	**#Miss**	
Our system	180	136	96	40	0
eSur @ trecvid 2010 SVM-HMM[15]	180	150	85	65	15
Classical SVM system [1]	180	121	54	67	40
Object Put Event					
Our system	180	136	120	16	0
eSur @ trecvid 2010 SVM-HMM[15]	180	140	85	55	15
Classical SVM system [1]	180	111	65	47	40
PersonRun Event					

[13] Q.Ye and al. Exciting event detection in broadcast soccer video with mid-level description and incremental learning. *Proceedings of the 13th annual ACM international conference on Multimedia*,pp.455-458,2005.

[14] National Institute of Standards and and Technology (NIST). TRECVID 2009 Evaluation for Surveillance Event Detection. ,2009.

[15] M.Takahashi, Y. Kawai, M. Fuji and M. Shibata. Semantic Indexing and Surveillance Event Detection. *TRECVID'2010*,2010.

[16] Wali and A. M.Alimi. Incremental Learning Approach for Events Detection from large Video dataset. *Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance*,2010.

[17] Wali and A. M.Alimi. Multimodal approach for video surveillance indexing and retrieval. *Journal of Intelligent Computing*,4:pp.165-175,2010.

[18] F. Wang, Y. Ma, H. Zhang and J. Li. A generic framework for semantic sports video analysis using dynamic bayesian networks. *Proc. of IMMC*,pp.115-122,2005.

[19] I.Laptev, M.Marszalek, C.Schmid, and B.Rozenfeld. Learning realistic human actions from movies. *In 2008 IEEE Computer Society Con- ference on Computer Vision and Pattern Recognition (CVPR)*,2008.

[20] G. Wang and Y. Cui. On line tool wear monitoring based on auto associative neural network. *Journal of Intel. Manuf.*,24:pp.1085-1094,2013.

[21] S. Yuan, G. Ping and W. Shu. Event detection: IPG-BJTU at Trecvid. ,2010.

[22] S. Z. Seyyedsalehi. Simultaneous Learning of Nonlinear Manifolds Based on the Bottleneck Neural Network. *Neural Processing Letters*,2013.

[23] H. Zhipeng, Y. Guangnan and J. Guochen. PKU@TRECVID2009: Single-Actor and Pair-Activity Event Detection in Surveillance Video. ,2009.

[24] Y. Aribi, A. Wali and A.M. Alimi. An intelligent system for video events detection. *9th International Conference on Information Assurance and Security (IAS)*,pp. 108113, 2013.

[25] M. Chakroun, A. Wali, Y. Aribi and A.M. Alimi. Video event detection using auto-associative neural network and incremental SVM models. *15th International Conference on Intelligent Systems Design and Applications(ISDA)*,2016.

[26] Seong G. Kong b, Donglin Jin, Shengzhe Li and Hakil Kim. Fast fire flame detection in surveillance video using logistic regression and temporal smoothing. *Fire Safety Journal*,2016.

Author Biographies

Mohamed Chakroun Assistant on Computer Sciences at ISIK, University of Kairouan. He received the Computer Engineer Diploma in 2006 and the Master Diploma in 2010 from the National Engineering School of Sfax (ENIS), Tunis (Tunisia).
He is currently pursuing the Doctor degree in the National Engineering School of Sfax. Nowadays, he is a research member in the Research Groups of Intelligent Machine (REGIM), (ENIS), Sfax (Tunisia).

Yassine Aribi Assistant on Computer Sciences at ISIMA, University of Monastir.Got his Ph.D. in Engineering Computer Systems at National school of Engineers of Sfax, in 2015. He is member of the REsearch Groups on Intelligent Machines (REGIM-Lab.) and head of the HR processes in the same laboratory. His research interests include Computer Vision and Image and video analysis. These research activities are centered around Medical images analysis. He is IEEE Graduate member.

Ali Wali Assistant Professor on Computer Sciences at ISIM, University of Sfax. Got his Ph.D. in Engineering Computer Systems at National school of Engineers of Sfax, in 2013. He is member of the REsearch Groups on Intelligent Machines (REGIM). His research interests include Computer Vision and Image and video analysis. These research activities are centered around Video Events Detection and Pattern Recognition

Adel M. Alimi He graduated in Electrical Engineering in 1990. He obtained a Ph.D. and then an HDR both in Electrical Computer Engineering in 1995 and 2000, respectively. He is full Professor in Electrical Engineering at the University of Sfax, ENIS, since 2006.