Thesis Subject: Learning People Dynamics in Video Understanding

Guido T. Pusiol

May 19, 2008

1 Introduction

Cameras are continually in use all around, to monitor traffic, for security in private and public places and even at our homes. Because of these huge volume of data, it is necessary to develop efficient methods for video management. It is almost an impossible task to continually monitor these sources manually. While the understanding of video in general seems a complex task, it can be successful when analyzing particular scenarios by incorporating substantial domain knowledge. This work aims to develop a general framework to analyze video in different domains. To perform such task the framework will be restricted to the scene events evidenced by the motion of objects.

The framework is going to be powered by the VSIP platform developed at INRIA's Pulsar-Team to discriminate the different type of objects keeping only those recognized as persons; also the tracking will be handled by the platform to detect trajectories.

Often the trajectories that appear in video data are not randomically but they hide underlying structures (e.g., roads, sidewalks). The proposed framework will focus on the detection and understanding of these structures to determinate human activities.

The framework aims at building systems which will be able to work on-line, detecting in real time normal and abnormal activities that occur in the scene.

The objective of this work is to study learning techniques to build this framework. To validate this framework we will develop applications (i.e., real-time automatic activity monitoring systems) to be applied in different domains.

2 The Framework

The framework is composed four independent stages: Learning, Optimization, Semantics Extraction and Evaluation. Given a set of trajectories the learning stage will cluster this trajectories to identify the hidden structures. This stage includes also updating methods to adapt on-line these clusters.

The optimization stage is needed to improve and tune dynamically the parameters used in the learning stage.
The semantics stage comprises the high level analysis that can be done with the trajectories structures to extract meaningful human activities.
Finally the evaluation stage describes the methods to validate the performance of all stages of the framework.

2.1 Learning

The goal of the learning stage is to cluster a set of given trajectories to determine the hidden structures of this set. These clusters are characterizing the most frequent human activities (e.g., interacting with an equipment) in the scene but also the topology of the scene environment (e.g., zones of interest such as stopping zones).

A first point should be to select the best features characterizing the trajectories, such as entry/exit point, and the best features characterizing the scene topology such as areas of low/high speed will be evaluated.

A second point is to design a similarity distance between trajectories and trajectory clusters. Usually these distances have many parameters that can be tuned depending on the application objectives and the scene type.

In previous work we can find several approaches to learn the spatial paths and the regions of interest [1] [2] [3] [4]. Other techniques to learn people dynamics (e.g., a running person that suddenly slows down and starts walking) include Hidden Markov Models [12] [13] which structures can be learned.

Taking in to account the state of the art, we will focus on the extension of the approach proposed by Patino *et al.* [7], where the representation of the trajectories is flexible and compact, and the clusters extraction approach is done by the measure of the Euclidean similarity distance of the trajectories features (i.e., 2D/3D coordinates at each point, duration, distance from origin to destination).

Although it is interesting to analyze complete tracks, also is important to recognize and evaluate a partial trajectory as it occurs to be able to predict where are people final destinations.

After the training procedure the framework should be able to classify new incoming trajectories into the trained model (i.e., the trajectory clusters), also determinate anomalous behaviors (detected trajectories that do not fit into the learned groups). Claudio Piciarelli's PhD. [9] thesis is directly related to anomalous behaviors detection, also Neil Johnsosn's thesis [18]. These approaches needs to be fully validated and cannot be applied as it is to unstructured scenes such as subway stations.

Since there is not guarantee that the learned models are stable in time, they should be incrementally updated during the on-line performance of the system. For example, the presence of a new obstacle in the middle of a learned path, will probably change the flow where the people walk. Moreover, trajectories that were detected as abnormal were probably detected as such because the frequency of appearance was low. In previous work we can find that Hu *et al.* [10] propose a batch update procedure for model addition, and Gales *et al.* [11] propose using maximum likelihood linear transformation, for an on-line fashion update. A difficult issue is to distinguish data representative to a new cluster from noisy data wrongly detected by the vision algorithms. This issue is crucial in video understanding because most of the data are corrupted by noise.

Thus the framework should extract the hidden structures, enable trajectory prediction, but also it will have an on-line mechanism to detect abnormalities

2.2 Optimization

Each trajectory is represented by features, used in the clusterization process as mentioned in the previous stage. As said before the learning stage contains many parameters to be tuned depending on the application objectives and scene type. These parameters include the selected features (and their weight) characterizing the trajectories, the parameters of the similarity distances and the parameters of the clustering algorithms. An important issue is to normalize and weight the features. For instance clustering algorithms has the tendency to rely more on binary features [7].Thus to achieve the domain independence a tuning stage is required [8].

The main work in this stage is to develop optimization mechanisms maximizing performance measures. In the state of the art many criteria have been proposed (e.g., Silhouette). A typical optimization mechanism (e.g., gradient descent or montecarlo) consists then in maximizing an energy function corresponding to one criterium. However all the criteria are not consistent (e.g., some need to be maximized and others minimized) and cannot be combined directly. To solve this issue, Genetic algorithms and Particle Swarm Optimization [5] using min-max strategy can be applied to maximize multiple performance measures used

to evaluate the quality of the clusters.

2.3 Semantics

Given the trajectory clusters, the semantic extraction stage consists in learning the topology of the scene environment by identifying the zones characterizing human activities. For instance, the scene contextual objects (e.g., a desk, an elevator, a machine), together with the regions of interests (e.g., a stop zone, forbidden areas) can be learned from the trajectory clusters and by analyzing specific spatio-temporal relationships. Ticket vending interacting zones can be learned by computing areas where people come, stop for a little while (queue if necessary) and then leave.
Zones of interest are not limited to trajectory properties. Other types of people dynamics can be explored. For instance a place where people stop while their size decreases can correspond to a sitting location.
Patino *et al.* [6] [7] and André [14] have defined the primitive events that occur in videos(i.e., inside zone, close to, stays_at, etc). The combination of the user defined objects, predefined primitive events with the learned scene topology can infer the semantics describing the observed scene.

2.4 Evaluation

The evaluation stage will validate the three previous stages of the framework. The quality of the proposed clustering approach will be validated by computing two main performance measures defined by Patino *et al.* [7], namely, Confusion and Dispersion. Also will be considered the performance indexes of traditional methods such as Davies-Bouldin [15], Silhouettes [17], and Dunn's [16].

The Confusion and Dispersion performance measures rely on ground-truth (i.e., the main routes corresponding to trajectory clusters). This ground-truth mostly contains starting and ending zones and needs to be extended to allow the validation of dynamic properties of the trajectories (e.g., the mean speed of the trajectory). Also ground-truth to validate unfinished trajectories prediction methods will be proposed.

Two main issues in this stage consist in taking into account very noisy data and to be able to handle routes with strong overlaps.

The final evaluation of the framework will be the field test. For this, the result application of this work will be taken into a Hospital. Thus the performance as a healthcare monitoring system will be evaluated in a real world environment.

References

[1] D. Markis and T. Ellis, *Learning semantics scene models from observing activities in visual surveillance*, IEEE Trans. Syst. Man, Cybern. B, vol. 35, no. 3, pp. 397-408, June 2005.

[2] D. Markis and T. Ellis, *Path Detection in Video Surveillance*, Image and Vision Computing, 20(12), pp. 895-903, October 2002.

[3] J. Owens and A. Hunter, *Application of self-organism map to trajectory classification*, in Proc. IEEE Visual Surveillance, July 2000, pp. 77-83.

[4] C. Staufer and W.E.L Grimson, *Learning patterns of activity using real time tracking*, IEEE Trans. Pattern Anal. Machine Intell., vol 22, no. pp.747-757, Aug 2000.

[5] Kennedy, J. and Eberhart, *R. C. Particle swarm optimization.* Proceedings of IEEE International Conference on Neural Networks, Piscataway, NJ. pp. 1942-1948, 1995.

[6] J. Patino, H. Benhadda, E. Corvee, F. Brmond, M. Thonnat., *Video-Data modeling and Discovery*4th IET International Conference on Visual Information Engineering VIE 2007, London, UK, 25th - 27th July 2007.

[7] J. Patino, H. Benhadda, E. Corvee, F. Brmond, M. Thonnat., *Extraction of Activity Patterns on Large Video Recordongs*, IET Computer Vision, accepted paper 2008.

[8] Anjum, N., Cavallero,A, *Single camera calibration for trajectory-based behaviour analysis*, IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, AVSS '07, 2007.

[9] Claudio Piciarelli, *Trajectory clustering techniques for unsupervised anomalous event detection.* Universita Degli Studi di Udine, Dipartimento di Matematica e Informatica, Italia.

[10] W.Hu, D.Xie, T.tan, and S.Maybank, *Learing activity patterns ussing fuzzy self-organizing neural networks*, IEEE Trans. Syst., Man, Cybern. B, vol. 34, no. 3, pp. 1618-1626.

[11] M. Gales, D.Pye, and P.Woodland, *Variance compensation within the MLLR framework for robust speech recognition and speaker aceptation*, in Proc. IEEE Intl. Conf. Spoken Language, Oct 1996, pp. 1832-1835.

[12] Porikli, F., *Learning object trajectory patterns by spectral clustering*, Proc. IEEE Int. Conf. on Multimedia and Expo ICMEv '04, 2004, 2, pp.1171-1174.

[13] Bashir, F. I., Khokhar, A. A., Schofeld, D., *Object Trajectory-Based Activity Classification and Recognition using Hidden Markov Models*, IEEE Transactions on Image Processing, 2007, 16, (7), pp. 1912-1919.

[14] E. André, G. Herzog, T. Rist, *Natural Language Access to Visual Data: Dealing with Space and Movement*, 1st Workshop on Logical Semantics of Time, Space and Movement in Natural Language, Toulouse, France, 1989.

[15] D.L. Davies, D.W. Bouldin, *A cluster separation mesuare*, IEEE Transactions on Pattern recognition and Machine Intelligence, vol 1, No. 2, pp. 224-227, 1979.

[16] J. Dunn, *Well separated clusters and optimal fuzzy partitions*, J. Cybernetics, vol. 4, pp. 95-104, 1974.

[17] P.J. Rousseeuw, *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, J. Comp App. Math, vol. 20, pp. 53-65. 1987.

[18] Neil Jonson, *Learning Object Behaviour Models*, Phd thesis, School of Computer Studies, University of Leeds, UK, September 1998.