Università degli Studi di Salerno

Dottorato di Ricerca in Ingegneria dell'Informazione
XII Ciclo – Nuova Serie

Doctorat de l'université de Caen Basse-Normandie

PHD THESIS

Detecting and indexing moving objects for Behavior Analysis by Video and Audio Interpretation

CANDIDATE: **ALESSIA SAGGESE**

SUPERVISORS: **PROF. MARIO VENTO**
PROF. LUC BRUN

COORDINATORS: **PROF. ANGELO MARCELLI**
PROF. FRÉDÉRIC JURIE

Academic Year 2012 – 2013

In the last decades we have assisted to a growing need for security in many public environments. According to a study recently conducted by the European Security Observatory, one half of the entire population is worried about the crime and requires the law enforcement to be protected.

This consideration has lead the proliferation of cameras and microphones, which represent a suitable solution for their relative low cost of maintenance, the possibility of installing them virtually everywhere and, finally, the capability of analysing more complex events. However, the main limitation of this traditional audio-video surveillance systems lies in the so called psychological overcharge issue of the human operators responsible for security, that causes a decrease in their capabilities to analyse raw data flows from multiple sources of multimedia information; indeed, as stated by a study conducted by Security Solutions magazine, after 12 minutes of continuous video monitoring, a guard will often miss up to 45% of screen activity. After 22 minutes of video, up to 95% is overlooked.

For the above mentioned reasons, it would be really useful to have available an intelligent surveillance system, able to provide images and video with a semantic interpretation, for trying to bridge the gap between their low-level representation in terms of pixels, and the high-level, natural language description that a human would give about them.

On the other hand, this kind of systems, able to automatically understand the events occurring in a scene, would be really useful in other application fields, mainly oriented to marketing purposes. Especially in the last years, a lot of business intelligent applications have been installed for assisting decision makers and for giving an organization's employees, partners and suppliers easy access to the information they need to effectively do their jobs.

The aim of this thesis is to face the above-mentioned issues, as fascinating as challenging since requiring different knowledge and skills, ranging from computer vision and pattern recognition to machine learning and databases. Indeed, there are a lot of issues, which make such a task very complex, and not yet solved in a definitive way; first of all, the system has to detect the objects of interest populating the scene (for instance persons, vehicles or animals). It is worth pointing out that this is a very complex task to achieve since, although being a moving object, a tree moved by the wind can not be considered an object of interest and then the system needs to filter out this kind of patterns.

One of the most challenging problem is related to the occlusions, which are due to the perspective flattening introduced by the use of a single camera: in the above mentioned systems it is required that the person is tracked (and then that its trajectory is extracted) also if it is partially or completely hidden by another person moving in the scene or by a static object. Think, as an example, to a wall which only allows to see a small part of the person (his head or his legs), making this task extremely demanding.

Another arduous task is the managing of the acquired trajectories: as a matter of fact, in real scenario it is required that billions of trajectories, extracted from different cameras, must be stored and that on this wide database the human operator can submit complex queries involving geometric and temporal data, as well as semantic one. Furthermore, each object trajectory needs to be properly analysed in order to interpret the behaviour and to understand if it refers to an abnormal one.

As for the design and implementation of a reliable and general audio event detector, one of the main open issue is that the properties characterizing the different events of interest might be evident at very diverse time scales: compare, for instance, an impulsive sound like a gunshot with a sustained sound, like a scream, that can have a duration of several seconds. Furthermore, in real applications there is often the problem that the sounds of interest are superimposed to a significant level of background sounds; thus it might be difficult to separate the noise to be ignored from the useful sounds to be recognized.

Finally, all this operations need to be performed in real time: it means that meanwhile the video and the audio are acquired (usually with a frame rate of 25 frames per second a bit rate of 128 Kbit per second) the system has to perform all the required operations: detection of the objects, tracking of their movements, storing and analysis of their trajectories, detection of audio events and finally allowing an interaction with the user. This last aspect in particular bestows to this thesis a high technological value, other than a significant scientific advance with respect to the state of the art.

In the following, more information about the proposed system are provided: it starts by analysing the videos and by extracting the trajectories of the objects populating the scene (tracking): it is important to underline that the trajectory is a very discriminant feature, since the movement of objects in a scene is not random, but instead have an underlying structure which can be exploited to build some models. The main novelties of the proposed tracking algorithm lie in the following aspects: first, the entire history of each object populating the scene is analysed by means of a Finite State Automaton; second, the updating of information related to each object is performed by a graph-based approach. Finally, the occlusions are properly managed by tracking into a different way single objects and groups of objects. The proposed tracking algorithm has been evaluated during an international competition (PETS 2013), ranking in the first places for all the considered scores over a high number of participants (more than thirty).

Once extracted, this large amount of trajectories needs to be indexed and properly stored in order to improve the overall performance of the system during the retrieving step: the main novelty of this module pertains the enhancement of off-the-shelf solutions, namely PostGis (the spatial extension of the traditional PostGres database) in order to deal with trajectories, which are very complex elements to manage for their spatio-temporal nature.

In general, the main advantage of the proposed approach lies in the fact that the human operator can interact with the system in different ways: first of all, he is informed by the system as soon as an abnormal behaviour occurs; abnormal behaviours can be associated to dangerous behaviours in the context of intelligent video surveillance, since referring to those objects drawing a trajectory in the scene which the system never saw. It is evident that the system has to be robust enough to deal with the errors typically occurring during the tracking phase, related for instance to broken trajectories. Furthermore, a high level of generalization is required in order to avoid wrongly classifying as abnormal those normal trajectories that only rarely occur. In order to cope with the above-mentioned issues, the similarity between trajectories is evaluated by means of a novel metric based on string kernel, especially defined for this purpose.

Whereas the information extracted from the videos are not sufficient or not sufficiently reliable, the proposed system in enriched by a module in charge of recognizing audio events, such as shoots, screams or broken glasses. It is worth pointing out that the integration between audio and video based information is a significant add-on for the proposed framework, being a completely novel aspect in the field of video and audio analysis.

In the proposed framework the human operator can also extract typical paths occurring inside a scene without any knowledge about the particular scenario (clustering): each path can represent, depending on the particular application domain, common interests of the customers in a shopping center or normal behaviours, related to both people moving into a train station and vehicles crossing an highway. The proposed clustering algorithm, based on a tree structure, exploits the similarity metric defined above in order to perform its task. This method has been evaluated over standard datasets, confirming its promising performance if compared with other state of the art approaches and its applicability in real and crowded scenarios.

On the other hand, the human operator can ask different typologies of queries to the proposed framework by personalizing the parameters only at query time, that is in the moment the query is thought: for instance, the objects crossing a given area in a given time interval (dynamic spatio-temporal query) or the objects following a particular trajectory, similar to the one hand drawn by the user (query by sketch).

Each proposed module has been tested both over standard datasets and in real environments; the promising obtained results confirm the advance with respect to the state of the art, as well as the applicability of the proposed method in real scenarios.