Multimedia analysis on user generated content for safety oriented applications

Nikolaos Papadakis[1], Antonios Litke[2], Anastasios Doulamis[3], Nikolaos Doulamis[3]

[1]Hellenic Army Academy
[2]Infili Technologies PC
[3]National Technical University of Athens

Paper abstract

The advancement of surveillance systems and early warning architectures is crucial for securing humans in large cities and prevent from illegal actions. This proves to be a key milestone to European economies, industries, authorities and at the end to the citizens' society. Nevertheless, the cost for large deployments and maintenance of ground sensing networks for local surveillance is extremely high, especially when the surveillance is operated under a manual environment. We, as humans, are subjectivity in interpreting the risk, may pay our attentions on different objects rather than on the real threat and most importantly we easily get tired especially in cases where we are obliged to monitor systems of multiple cameras with the chance of detecting a trespassing or an improper citizen's behavior.

The recent advantages in hardware and software technologies have significantly reduce the cost of visual sensors and they have forced the development of new innovative automated (or even semi-automated) algorithms that are able to localize on images objects/regions of interest so as to alert for a potential risk. Nowadays, there exists several surveillance cameras in cities both for security and preventing the crime but also for traffic management and control. On the other hand, the Internet, the new digital technologies and the social services, like Facebook and Twitter are transforming our world – in every walk of life and in every line of business. Nowadays, several trillions of images exist in loosely structured repositories (e.g. the Web, file servers, databases etc.) and their number grows rapidly every day. This is the natural outcome of a series of reasons such as the low-cost of digital cameras, the low-cost storage and easy Web hosting and on top of that the need of people and organizations to share multimedia files, either for social or commercial purposes. This huge amount of media information presents an opportunity for the computer vision and multimedia society to "perform automatic or semi-automatic" analyses of the content and localize object of interest.

This is the key goal of this paper to incorporate image analysis methods towards a more safe and secure environment for the citizens of a large-scale city. Towards this direction, initially we need to process, evaluate and prioritize video and audio streams generated by users (where users act as "sensors") in participatory urbanism approach. The proposed method prioritizes the multimedia streams based on their content (e.g. streams that involve situations where a critical situation is encountered such as panic of crowd, on-going street riots or other dangerous situations) enabling thus the authorities to better shortlist the actual streams and identify the events.
This is achieved by the application of a content-based filtering method which exploits (a) geo-reference information (if this is available) and (b) identifying scenes within the stream that potentially fall into the previous descriptions. The proposed solution is being studied in the framework of the City.Risks project and is based on the Red5 media server. Red5 is an open source media server for live streaming solutions of all kinds. It is designed to be flexible with a simple plugin architecture that allows for customization of virtually any Video On Demand and live streaming scenario. The user acting as a sensor is using his personal mobile device equipped with an Android application that is able to stream directly to the Red5 server.

Having discarding the irrelevant image content, we apply image processing and analysis algorithms to localize objects of interest. Towards this, machine learning tools are investigated and properly interwoven in the analysis tools. Today, the traditional machine learning and computer vision techniques exploit "shallow architectures" in learning. Although the conventional approaches assume a highly non-linear and adaptable model framework, such as conventional hidden Markov models, neural networks with adaptive capabilities,

linear or non-linear dynamic systems, maximum entropy models, kernel regression methods, support vector machines and conditional random fields, their common property is that they rely on a single layer of processing. Thus, the classification process is inherently limited in performance. On the contrary, the emulation of the efficiency and robustness by which the human brain represents information resembles a deep machine learning framework; humans' brains do not work by explicitly pre-processing sensory signals but rather allow them to propagate into complex hierarchies. This discovery motivated a new area in machine learning research with applications in image analysis problems; the deep learning paradigm, which focuses on computational models for information representation that exhibits similar characteristics to that of the humans. Another research challenge that we apply is the outcome of a detailed 3D map of a region based on relevant images. 3D (geometric) scene reconstruction is an important element for securing citizens since it allows a details numerical description of the geometry of a scene. Such a description also allows for a geometric modelling.