Video abstraction and detection of anomalies by tracking movements

Paolo Buono
Università di Bari
Dipartimento di Informatica
Via E. Orabona, 4, 70125 Bari
+390805442239
buono@di.uniba.it

Adalberto L. Simeone
Università di Bari
Dipartimento di Informatica
Via E. Orabona, 4, 70125 Bari
+390805442299
simeone@di.uniba.it

ABSTRACT
The increasing adoption of video surveillance makes it possible to watch over sensitive areas and identify people responsible for damage, theft and violence. However, when such events are not detected immediately, the subsequent video analysis can be a long and tedious task. The aim of this paper is to present a technique that allows a human investigator to focus only on those parts of a video showing the event as it unfolds, and so helping to save on the time needed to identify and understand how it happened. The presented technique creates a single interactive image of the whole video that shows everything that happened in the scene. The human investigator can then select an area of interest and those parts of the video related to that specific area will start to play.

Categories and Subject Descriptors
H.5 [**Information Interfaces and Presentation**]: *Multimedia Information Systems*

General Terms
Algorithms, Performance, Security, Human Factors.

Keywords
Video analysis, video abstraction, visual analytics

1. INTRODUCTION
In recent years, the field of video analysis and video surveillance has become even more important due to the wide popular attention paid to world events, which has exposed the shortcomings of the current state of technology. Video-based analysis focuses either on real-time threat detection or on recording video for subsequent forensic investigations.

A well defined network of surveillance cameras is often present in cities, aimed at maintaining a comprehensive coverage of specific locations such as a building or other sensitive areas. The main goal of these systems is to assist users to detect and identify potential threats or suspicious events arising during the recorded timeframe of the video.

AVI '10, May 25-29, 2010, Rome, Italy

It has been demonstrated that the human attention span can drop below acceptable levels very quickly (after only 20 minutes), even in trained observers [2]. As a consequence, video monitoring can be very ineffective. Contributions from realtime algorithms, dealing with object identification, detection and tracking, are therefore needed to alert surveillance staff when suspicious events occur.

This work addresses the analysis of video produced by stationary surveillance cameras such as those typically placed in outdoor urban settings. This technique is not intended for alerting people while the event occurs; it is intended for post-processing activities. The analysis is done after the event has occurred and the main goal is to reduce the time spent for video analysis.

Current technology does not allow a purely automatic video interpretation that can provide answers to a wide set of possible questions. Some events may be highly unpredictable. For effective video analysis it is important to combine the perception, flexibility, creativity and general knowledge of the human mind with the enormous storage capacity and computational power of computers. This activity is often very time-consuming, and the whole process needs to be speedier so that the analyst can focus her/his efforts on the relevant parts of the video without wasting time on meaningless segments.

In the next section, an overview of related work is provided. In Section 3, the video analysis method is presented, while in Section 4 some of the test scenarios are described. Finally, conclusions and possible future research directions are discussed in Section 5.

2. RELATED WORK
A comprehensive review of the state of the art in automated video surveillance technologies can be found in [8]. In this work, the authors analyze the state of the art in this field with reference to several aspects: image processing, surveillance systems and system design. One of the most prominent of these is certainly the issue of computer vision algorithms. Another issue is related to the possible ways in which to integrate these different algorithms and approaches to build a complete surveillance system. Finally, the last issues concerns large scale surveillance systems, such as those present in public transportation systems (subways, airports, etc.) or large buildings where there is a high concentration of people in small spaces.

User interfaces allowing users to operate these detection systems also play a key role. From this point of view, a ”surveillance index browser”, part of a larger architecture that also comprises a face-recognizer module, is presented in [3]. The user interface shows, on a timeline, an overview of all the events detected by the system in a particular time-frame. A second timeline provides a zoomed version of any video segment chosen by the user. A window displays the output of the tracker camera, while a second window displays a zoomed-in video feed of the moving object.

Huston et al. present a system whose key feature is ”the ability to perform user-defined queries on unstructured surveillance data” [4]. The rationale lies in taking advantage of the computer resources for searching video data, allowing the user to focus her/his attention on interpreting the results in the hope of gaining insights that might help to accomplish the task s/he is undertaking. For example, in the case of forensic investigations, thanks to the user interface provided, the user may select a portion of an image (e.g.: a person, a suitcase, a car, etc.) where something suspicious is happening (e.g.: a video frame showing a bank robber leaping into a getaway car) and then start the search. The application utilizes a brute-force search algorithm for performance reasons. The most important of these is that it is generally not expected to know what to search for, so indexing criteria are useless. The authors point out that since the rate at which new data is generated far exceeds the rate at which it can be analyzed and most of it will never be searched before being erased, it follows that preprocessing is a waste of resources. Application-specific code starts the search at each physical storage device instead of at a centralized place. Users may also fine-tune the search by modifying the algorithm parameters. The application supports a number of search parameters such as color, shape, texture and object detection. Once a search is started, the user may continue her/his analysis of the video. An advantage of this system is that partial results are shown as soon as they are acquired. If one of these matches shows promising results, the user may also start concurrent searches while waiting for others to be completed, or stop unpromising leads altogether.

DOTS [1] is an indoor multi-camera surveillance system for use in an office setting. The user interface displays an overview of the viewable cameras through a series of small thumbnails. With these cameras a person can be tracked and his/her position estimated by mapping the person’s foreground shapes on a 3D model of the office. In this process, the changing position of a person can be tracked across multiple cameras. A floor plan displays the current position of the employees in a given instant. The currently tracked person is shown in a bigger preview area in the interface, complete with the person’s detected face. This system’s main advantage is its employment in a known and stationary setting, although this can also be seen as the main disadvantage.

An original approach by Vural et al. [9] uses eye-gaze analysis [6] to extrapolate, by means of video abstraction techniques, a processed video which is an integration of the video parts showing those actions the operator is focusing on or overlooking. In this way, relevant parts of the video can be efficiently and rapidly reviewed, without having to go through the whole video.

A similar approach to our own is used in [7] where ”slit-tear visualizations” are drawn on the source video. For every frame the system writes on a timeline the pixels beneath these lines. For example, if such a line is drawn on the side of a road, crossing the resulting timeline the shapes of the car passing along it will be shown. These shapes will be more or less elongated depending on the speed. If nothing is happening, the line will keep writing the same pixels. So any event that happens over that line will be easily identifiable because the background pixels, being static, will be uniform, so foreground objects will stand out.

3. THE METHOD

Our approach relies on the computation of an image which summarizes the activity happening in a scene. This process displays traces of movements across the scene. The tool works best with a stationary camera and is very useful when it is known what happened, but the precise details about the how and when are unknown (i.e. a bicycle is stolen but it is not known when and from who) . The image uses a color code scale ranging from red to yellow to show how old the events are. In order to highlight events that occurred in a brief time, the user can choose to use a logarithmic scale. The human investigator can browse and review the segments of video which depict the events of interest by clicking on areas of interest. From an algorithmic point of view, the video analysis and image generation process consist of the following steps:

1. data structure initialization;
2. movement analysis;
3. noise reduction;
4. sequences matrix loading;

Initially, an N x M matrix, shown in Figure 1 and called sequences matrix is created. The matrix contains information about any persons or objects that moved in the video. N and M are the rows and columns of the original video frame resolution, respectively. The first element, which is the most frequently accessed, stores the total amount of points or objects the sequence is made of. The next four elements are pointers to the other neighboring sequences (both horizontal and vertical).

The video is then analyzed using the frame difference algorithm, an algorithm employed for the purpose of detecting the movement of objects across two (single difference) or three frames (double difference [5]). This algorithm generates a new video containing only information about such movements. The technique addresses also cameras that rotate regularly to several predefined positions. During rotation, all frames representing the camera movement are discarded. Scene segmentation algorithms are applied in order to isolate video segments having a fixed orientation. The segmentation is based on a threshold level, and once this is exceeded the scene is assumed to have changed. The algorithm presented here analyzes each scene separately and a sequence matrix for each scene is loaded.

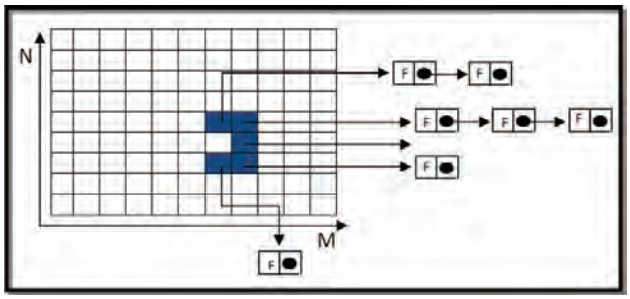

Figure 1. N x M matrix representing pixels in which movements occurred

Each "cell" of the sequences matrix identifies a pointer to an OpenCV Sequence, whose elements are integers representing the frame number in which a movement was registered. The video starts from frame 1 and proceeds sequentially. If no movement is detected in the whole video in a specific pixel, the corresponding sequence will be empty. The first element of the cell, if not empty, contains the length of the corresponding sequence, which represents the amount of activity for the associated pixel.

The matrix is used to obtain an image, by assigning to each value a different color on a red-yellow "temporal" gradient: pixels showing a red tint represent activities that occurred further back in time, while pixels with a yellow tint represent more recent activity. The resulting image shows all the movements that occurred in that given scene; an example is shown in Figure 2 (a).

Figure 2. Screenshots taken from the analysis of the bike theft scenario

4. SCENARIOS

In a typical ordinary scenario, some adverse event will have happened: someone may have stolen a valuable object or vandalized a store and so on. The human investigator will be presented with the "fait accompli" but no clue whatsoever about the specific time and how the event unfolded. If no other technological devices were available (like motion detectors) the only option left would be to review hours of video recordings in the hope of identifying the exact moment related to the event. This is obviously very time consuming and, as previously noted, the watcher is prone to losses of attention span. The tool streamlines these activities by presenting the user with an abstraction of the movements registered in the whole scene which is overlaid on the background. Only persons or objects which moved leave a trace on the resulting image.

This technique has been employed in several scenarios to evaluate its effectiveness. In the following one, a video camera is focused from above on a small crossroad. In Figure 2, a cyclist passes by (b) and leaves his bike on a nearby wall (c). After some time another person enters the area, steals the bike (c) and rides away. The resulting abstract image (a) clearly shows two differently colored movement traces: the red trace is generated by the movement of the cyclist who leaves the bike, while the yellow trace is generated by the person who steals the bike.

With this tool, the human investigator can draw a rectangular-shaped selection area over the image. In this case, the best solution is to draw the selection at the point where the two traces cross (where the bike was left); at this stage a segment with all the frames inside the defined area where movement was registered will start playing. In our example, after clicking, the investigator will be shown the part of the video where the cyclist leaves the bike and then another part where the thief steals the bike and rides away.

If more information about the event is known, as in the case of a theft in a store, the process is even more streamlined. Let us consider the example of a fixed camera in a small convenience store: here an object has gone missing from the counter in the bottom left corner of the recorded scene. The output of the system shows a very noisy result due to all the people walking to and in the store (Figure 3 (a)). But, by knowing the location of the missing object, it is possible to isolate only those video segments that show people interacting with the dispenser. Thanks to this feature, the culprit of the theft is quickly identified (b). The operator only needs to select the area comprising the dispenser and review potential suspects interacting with the dispenser until he finds the segment showing the thief stealing the missing object. The amount of work is greatly reduced because hours of unhelpful video segments are removed leaving only those that can contribute to the identification of the thief.

Figure 3. (a) shows the system output when applied in a crowded store, (b) shows the video segment obtained by isolating the movement detected in a selected area of the scene

5. CONCLUSIONS AND FUTURE WORK

In this paper, a technique for detecting relevant moments in a video recording is presented. The technique allows human investigators to interact with a single image that reveals what happened in a scene taken by a stationary video surveillance camera. One of the advantages of this technique is its general purpose applicability: in fact, it has been tested and proved effective in outdoor and indoor settings.

Thanks to the use of this technique, users were able to reconstruct the story in a very short time, rapidly pinpointing the object of the search, even if this was not known beforehand. This exploratory use of the system revealed some limitations, which will be addressed in a future work.

There is room for improving the detection algorithm by incorporating better object detection and background/lighting filtering. In fact, the main objective of this preliminary work was to demonstrate the feasibility of this technique and its effectiveness in obtaining good results in a variety of situations. A more satisfactory user interface could be designed to better support analysts. The first step will be to allow for greater flexibility when selecting the area of the video frame to be checked for movements: as the system is still in its prototypal step the extent of this area is fixed. Other improvements that are currently under study are ways for the system to better convey the most potentially important events of interest in the image and the possibility of comparing several video segments at once.

6. REFERENCES

[1] Girgensohn, A., Kimber, D., Vaughan, J., Yang, T., Shipman, F., Turner, T., Rieffel, E., Wilcox, L., Chen, L. and Dunnigan, T. 2007. DOTS: support for effective video surveillance. In Proceedings of the 15th International conference on Multimedia (Augsburg, Germany, September 24 - 27, 2007). Multimedia '07. ACM Press, New York, NY, USA, 423–432

[2] Green, M. W. 1999. The appropriate and effective use of security technologies in U.S. schools. A guide for schools and law enforcement agencies. Technical report, Sandia National Labs, Albuquerque, NM, USA.

[3] Hampapur, A., Brown, L., Connell, J., Ekin, A., Haas, N., Lu, M., Merkl, H. and Pankanti, S. 2005. Smart video surveillance: exploring the concept of multiscale spatiotemporal tracking. IEEE Signal Proc. Mag. 22,2 (March 2005), 38–51.

[4] Huston, R. Sukthankar, J. Campbell, and Pillai P. 2004. Forensic video reconstruction. In Proceedings of the ACM 2nd International workshop on Video surveillance & sensor networks (New York, NY, USA, October 15, 2004). VSSN '04. ACM Press, New York, NY, USA, 20-28.

[5] Kameda Y. and Minoh, M. 1996. A human motion estimation method using 3-successive video frames. In Proceedings of the International Conference on Virtual Systems and Multimedia (Gifu, Japan, September 18 – 20, 1996). VSMM'96. 135–140.

[6] Pritch, Y., Rav-Acha, A., Gutman, A. and Peleg, S. 2007. Webcam synopsis: Peeking around the world. In Proceedings of the IEEE 11th International Conference on Computer Vision (Rio de Janeiro, Brazil, October 14 - 20, 2007). ICCV 2007. IEEE Computer Society, Los Alamitos, CA, USA, 1-8.

[7] Tang, A., Greenberg, S. and Fels, S. 2008. Exploring video streams using slit-tear visualizations. In Proceedings of the working conference on Advanced Visual Interfaces (Napoli, Italy, May 28 – 30, 2008). AVI '08. ACM Press, New York, NY, USA, 191–198.

[8] Valera, M. and Velastin, S. 2005. Intelligent distributed surveillance systems: a review. IEE Proc., Vis. Image Signal Process. 152,2 (April 2005), 192–204.

[9] Vural, U. and Akgul, Y. 2009. Eye-gaze based real-time surveillance video synopsis. Pattern Recognition Letters, 30,12 (September 2009), 1151–1159.