User-Interface to a CCTV Video Search System

Hyowon Lee, Alan F. Smeaton, Noel O'Connor, Noel Murphy

Centre for Digital Video Processing, Adaptive Information Cluster, Dublin City University, Ireland

Abstract. The proliferation of CCTV surveillance systems creates a problem of how to effectively navigate and search the resulting video archive, in a variety of security scenarios. We are concerned here with a situation where a searcher must locate all occurrences of a given person or object within a specified timeframe and with constraints on which camera(s) footage is valid to search. Conventional approaches based on browsing time/camera based combinations are inadequate. We advocate using automatically detected video objects as a basis for search, linking and browsing. In this paper we present a system under development based on users interacting with detected video objects. We outline the suite of technologies needed to achieve such a system and for each we describe where we are in terms of realizing those technologies. We also present a system interface to this system, designed with user needs and user tasks in mind.

1. INTRODUCTION

More and more surveillance CCTV cameras are being deployed in factories, company buildings, subways and on the streets for real-time monitoring and as well as for recording for forensic evidence and analysis. These cameras usually record 24 hours a day, everyday, resulting in a very large volume of stored video data. Most digital CCTV systems allow security staff to view recorded video contents using playback, fast-forwarding and rewinding features. In order to retrieve a particular video segment following an incident, knowing the date/time and the location of the incident is essential because of the time and effort that it would take to manually look at all the recorded video. Even with knowledge of the time and location, looking for interesting or suspicious events close to an incident in the recorded CCTV video archive requires considerable effort from security staff who do searching, by fast-forwarding until something interesting is found and then watching for a person, people or objects that draws attention. Having located a person or an object in the video, trying to find that same person or object from other videos taken from a nearby camera is an even more difficult task because the time when the person or object could be located can be unpredictable. Finding *all* instances of a given person or object in a given day from all cameras on a campus for example would be indeed a very difficult task to do manually, if it was possible at all, though this would be an extremely useful feature to help determine what events happened.

Many areas of research and development can be applied to help more effective retrieval of recorded CCTV videos. Currently work is underway in the areas of hardware (camera, sensor management), system architecture (concurrency, distributed networks), and content-based video retrieval (object detection, segmentation, tracking, classification) all contributing to developing a new generation of CCTV surveillance systems and tools that could reliably analyse content in real-time and generate an alert when a specified event happens. Additionally we require support for forensic analysis through complex queries and visualisation of events from the network of cameras. This can be seen, for example, in papers from the 1st and 2nd ACM SIGMM Workshops on Video Surveillance and Sensor Networks over the past two years (1, 2).

However, as is the case in many other technology-related fields, the *user-interface* and *usability* for a system that could allow its users to efficiently and effectively search, browse and relate events from CCTV video content have not been yet studied in any depth. Without realising it, we might fall into a scenario in which all the advanced and promising underlying technologies end up mismatching the end users' searching and browsing needs. The poor usability of systems resulting from a technology-driven approach fail to satisfy even a minimum level for them to be usefully used. To illustrate, not a single paper in the aforementioned ACM SIGMM Workshops addressed the user-interface in their reported surveillance systems! As experienced again and again in other application areas of technology, user-interface and usability concerns are addressed much later in the CCTV development process, costing much and delaying further the appearance of operational, usable systems which bring the underlying functionality of the system to its maximum usability.

According to conventional HCI (Human-Computer Interaction) practise, system development starts with obtaining user and usability requirements leading to early prototyping to picture what the final system should be like before showing it to the potential users for opinion and feedback. This iterative refinement process is carried out before the underlying technology becomes assembled and complete. This is now a well-known principle in the system development process that has been advocated for decades in order to save development cost and develop systems that meet end-users real needs.

In the Centre for Digital Video Processing in Dublin City University (http://www.cdvp.dcu.ie), we are researching content-based video indexing and retrieval for various usage scenarios and application areas.

Currently one of the major areas of our research is addressing object-based processing in video covering tasks such as object segmentation, tracking, and the subsequent searching and browsing. One of the application areas for this stream of research is the development of an object-based CCTV surveillance system. In this work we see our system prototyping approach being usefully adapted to define possible user-interaction scenarios for such a system.

In this paper, we describe our approach to developing an object-based CCTV video search system. We start by sketching a possible user-interface based on the interaction style that appears in the video retrieval community to provide interactive search/browse tools for digital video retrieval systems. Following that we can then consider the kinds of underlying mechanisms that we have or will have in the near future which would turn the sketched interface into a reality. In this way, the focus remains on the user and on usability while the individual components of the underlying technology are evaluated separately and drawn together to realise the final system.

The paper is organised as following: in Section 2, we describe a possible scenario in which a forensic analysis is required following some event, and we then present a user-interface and interaction scheme for an object-based CCTV video search system which allows a searcher to easily filter, search, browse, play and re-query only the worthwhile video segments to quickly locate interesting or suspicious persons or objects and help to understand the circumstances of the event by automatically linking between the same person/objects from other cameras. In Section 3, we enlist and briefly describe our research areas in content-based video indexing and retrieval that are or could be used to realise this user-interaction scheme. Section 4 concludes the paper with the current status and our plans in this work.

2. INTERACTION DESIGN

Consider the following scenario: a student's laptop computer is stolen from one of the computer labs on a campus. She reports this to campus security and mentions the rough time and exact date when the theft probably happened. The security staff now need to do a "forensic analysis", going through recorded video footage from the CCTV cameras covering the two entrances to the lab in question around that time. An example of the user-interface to a typical contemporary system, operational within our campus and developed by us, is shown in Figure 1.

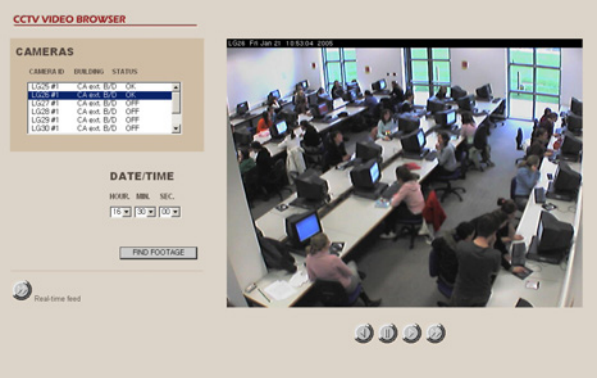

Figure 1. Simple CCTV search interface

The simple interface in Figure 1 allows security staff to select a camera and time, allowing access to the last 24 hours of CCTV video in/around computer labs on campus. Using this example system and indeed for any of the currently available CCTV surveillance systems, security staff would search the video archive by time and date and play, fast-forward, or rewind to locate suspicious events from multiple camera sources to try put together the circumstance of the incident, which could take hours, if not days.

With our new CCTV archival system, security staff are able to easily find out exactly what happened in this incident by using a search tool that automatically detects, correlates and visualises persons and objects at various times and multiple locations in the video archive. Without thinking too much of the technical implications of this for now, let us consider some potentially useful user-interface features for security staff in such a situation:

- Filtering out footage segments in which nothing happens at all – putting aside possibly a large amount of obviously unimportant footage allows security staff to focus only on footage where something happens;
- Presentation based on keyframes to capture the event – each event can be presented as a single keyframe that captures the essence of the event, thus eliminating the need for constant fast-forward and rewind actions which are time-consuming;
- Locating a suspicious person for subsequent querying in order to find all instances of that particular person during other times in the camera's recording, or from all cameras on a given day, and
- Visualising a person's whereabouts throughout a given day to allow easy tracing of the person's route, places s/he visited and duration of the visit at each location

Once an "event" – a segment of a video footage in which something happens – is defined, detected and indexed as the retrieval system's "document" (a unit of retrieval), the system becomes a search tool through an "event" document database that allows its users to search, browse and link among these units similar to

the way many modern information retrieval systems feature as their interface. From this point, the user-interface to such a system can also be designed in the style of a document retrieval system focusing on meaningful interaction sequences and useful features for the users, rather than the typical multi-camera playback interfaces available today which combine real-time surveillance features with archive search features.

With this in mind, we have designed an interaction scheme for a CCTV archival search system on a university campus setting for about 150 CCTV cameras distributed throughout the campus. The interface is shown in Figure 2.

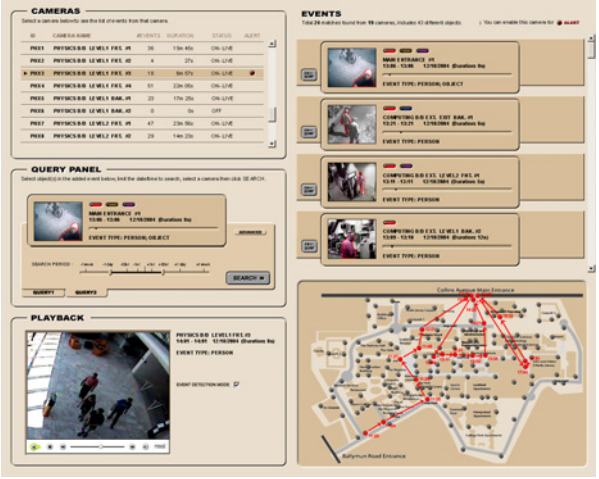

Figure 2. Overall interface

The interface is designed as a search/browse tool to support security staff in quickly finding out and tracing a suspicious person or object from any camera footage recorded during a given period of archival time. On the top left of the screen, the user sees a full list of cameras and on each entry the location, the count of events detected and the status of the camera are indicated. When a user selects one of the cameras, a summarised list of detected events from that camera is presented on the top right side of the screen. This event list means that all unimportant sequences where nothing actually happened have been filtered out, only displaying the footage worth inspection by an end user. Each entry represents an event in which one or more than one object has been detected, and is shown in more detail in Figure 3.

Figure 3. An event representation and interaction

The event representation in Figure 3 shows a number of information fragments, including:

- A keyframe taken from the middle of the sequence (the duration of which is 8 seconds in Figure 3);
- A set of buttons each of which represents a detected person or object found in the keyframe. Moving the mouse cursor over a button will highlight the button and the object within the keyframe that the button represents. In this way, the user's interaction with an object/person in the video is enabled;
- Camera ID;
- Date/time and duration of the event;
- Indication of this event's relative time within the archive on the timeline, and
- Type of the event (whether it is person or object that caused this event to be detected)

The user can browse the list of these event representations quickly using the scroll bar, and when locating a suspicious event can click on the keyframe to play the event in the playback panel (bottom left of Figure 2).

When the user succeeds in locating a suspicious person or object, he can use that person or object to query for other instances of this person or object detected as an event in other parts of this camera's footage, or in a number of other cameras near the location of this camera, or anywhere in the entire archive. Clicking on the QUERY button on the left of an event representation (left side of Figure 3) will copy the whole event representation into the query panel (left side of Figure 2). The user can then specify the time zone for this person/object occurrence by dragging the timeline widget below the copied event representation (for example, specifying the search range to between 1 day before and after this event), or use the ADVANCED button to specify the exact search scope with a start/end date and time and the cameras he wants to search. The default geographic search range is "nearby cameras" to the selected camera from which the copied event came. Clicking on the SEARCH button will trigger the retrieval of the selected person/object within the specified geographic (cameras) and chronological (time/date) range and the result will be displayed on the top right side of the screen.

The retrieval result is another list of event representations in each of which at least one of the detected persons or objects is the same as the one copied and specified in the query panel. The user can look at a map representation of the campus (bottom right of Figure 2) in which the matched person/object's trace is drawn as a red line, each point corresponding to one of the events represented in the search result list. Many of falsely detected persons/objects due to imperfect detection effectiveness (caused by changing lighting condition, skewed angles, occlusion, low resolution of the camera, etc.) can be filtered out and removed from the map based on simple temporal-geographic logic (for example, the same person cannot

exist in two different places at the same time or within a very short timeframe). The user can modify the search range in the query panel then click again on the SEARCH button to see a changed event representation list and the re-drawn route on the campus map. The user can also add any other event representation encountered during the interaction into the query panel for additional person/object examples in subsequent querying, or remove a previous query event from the query. Continuing to add and remove query events and browsing different search results is the iterative search process the user experiences with this interface.

Using this interface, the user browses and searches, refines the search using some of the suspicious persons/objects located during browsing and searching, zooming in to what footage seems interesting or suspicious. As can be seen, the user interface and the user interaction are more similar to recent experimental image/video retrieval systems, for example those presented in the annual series of TRECVID Workshops (3).

3. IMPLEMENTATION REQUIREMENTS

Having outlined the interaction scheme in the previous section, in this section we consider implementation - what kinds of underlying technologies in video analysis are required in order to realise a search tool such as this?

For several years we have been researching advanced video content analysis and applications which can be plugged into the system sketched earlier. In this section, we summarise our research streams which can be applied to realising the surveillance video search system for which the user-interface has already been designed, some of which are already reliable enough for deployment, some of which still require further work to bring their accuracy to an acceptable level:

- **Object segmentation and tracking** – the "holy grail" of video analysis is the accurate segmentation and tracking of objects in natural video. Research in this field over the past decade has resulted in reliable analysis on low-level features such as colour, shape and texture. However, aggregating these analyses to reliably detect semantic elements such as an object, has turned out to be a difficult task due to objects appearing at different angles when viewed with different cameras, problems with occlusion, lighting conditions (thus changing the colour and brightness of the object), and other numerous factors that make it difficult to automatically define and segment an "object." We have tackled this problem in different ways including the automatic detection of cartoon characters in the Simpsons TV programmes based on using characters' faces for template matching (4); we have developed semi-automatic segmentation in which an iterative refinement of an object outline is done using human intervention (5); we have explored the use of low-level features such as colour and texture as relevance feedback from a user to allow interactive object classification in order to achieve more accurate object search results; in the QIMERA and SCHEMA projects we have contributed to the development of a shared platform and environment for object detection (5, 6). We are also working on more reliable object segmentation with the use of an infrared video camera image overlaid on a visible spectrum camera image and we achieve object segmentation from the fusion of the two video sequences (7).
- **Pedestrian detection and tracking** – we have been experimenting with detecting and tracking pedestrians on a busy crossroads using 3D modelling with trinocular cameras and automatically removing backgrounds such as shadows, uneven ground plane noise, etc. (8).
- **Event detection** – we are working on the automatic detection of action and dialogue scenes in feature films using state machines for a set of low-level visual features such as motion and shot length combined with heuristic rules in film taking (9). In the context of surveillance video analysis an "event" can be based on the appearance of an object (i.e., from the entrance of an object to its exit from the field of view of the fixed camera).
- **Linking among similar objects** – we have been successfully applying the functionality of linking among similar or related documents in a variety contexts and applications and observing its importance in providing users with an effective and useful searching/browsing experience. In the Físchlár-News application, more than 7,600 news stories from daily TV news have been indexed and similarity among those stories determined from their closed caption analysis. This allows this system to automatically present "related stories" when a user is browsing a given news story (10).
- **Similarity among video keyframes** – automatically calculated keyframe similarity has also been applied in our interactive video retrieval systems which allow users to use any keyframes encountered during searching and browsing as a basis for subsequent querying to find similar keyframes in the video archives (11) and these have been based on content analysis of the keyframes using low-level features such as global and local colour, texture and edges, as well as simple semantic features such as the existence of a face in the keyframe or the fact that the shot is taken outdoors. Calculating similarity among objects detected in video will require something more than low-level features because an object's shape, size and colour can

appear very differently depending on the angle, distance and the lighting conditions in which it was taken.

4. CONCLUSION

We are currently at the stage where individual object detection algorithms we have been working on can be applied to CCTV videos. Some elements (object detection and tracking) are readily available, although without the accuracy we would like, as this field is still not mature enough for real applications yet. However, to demonstrate an integrated solution we are building a system with the designed user-interface as a front-end and many of the components mentioned above. Some elements (such as the use of trinocular cameras for pedestrian detection) are still at an early stage of development, but as we experiment further, these will be plugged into our overall system.

While the underlying techniques are being investigated and refined, the designed front-end interface will be also refined within its current prototype format; discussions and feedback from our campus security staff and subsequent iterative modifications will follow.

In our work on an object-based CCTV video retrieval system, we work from both ends of the development, namely from the individual components of the underlying detection and tracking methods, and also from the user-interaction and usability direction. The strength of this bi-directional development will show when these two sides eventually meet at the point of a full implementation.

Acknowledgments. Part of this material is based on works supported by Science Foundation Ireland under Grant No. 03/IN.3/I361. The support of the Informatics Directorate of Enterprise Ireland is gratefully acknowledged.

REFERENCES

1. 1st ACM SIGMM Intl. Workshop on Video Surveillance (IWVS'03), Berkeley, CA, 2003.

2. 2nd ACM Intl.Workshop on Video Surveillance and Sensor Networks(VSSN'04), New York, NY, 2004.

3. TRECVID: TREC Video Retrieval Evaluation. Website available at: http://www-nlpir.nist.gov/projects/t01v/t01v.html

4. Browne, P. and Smeaton, A.F. 2004. Video information retrieval using objects and ostensive relevance feedback. ACM Symposium on Applied Computing (SAC'04).

5. O'Connor, N., Adamek, T., Sav, S., Murphy, N. and Marlow, S. 2003. QIMERA: A software platform for video object segmentation and tracking. 4th European Workshop on Image Analysis for Multimedia Interactive Service (WIAMIS'03).

6. Izquierdo, E., Casas, J., Leonardi, R., Migliorati, P., O'Connor, N., Kompatsiaris, I. and Strintzis, M. 2003. Advanced Content-Based Semantic Scene Analysis and Information Retrieval: SCHEMA Project. 4th European Workshop on Image Analysis for Multimedia Interactive Service (WIAMIS'03).

7. O Conaire, C., Cooke, E., O'Connor, N., Murphy, N. and Smeaton, A.F. 2005. Fusion of infrared and visible spectrum video for indoor surveillance. 6th Intl. Workshop on Image Analysis for Multimedia Interactive Service (WIAMIS'05).

8. Kelly, P., Beardsley, P., Cooke, E., O'Connor, N. and Smeaton, A.F. 2004. Detecting Shadows and Low-lying Objects in Indoor and Outdoor Scenes Using Homographies. IEE Intl. Conference on Visual Information Engineering (VIE'05).

9. Lehane, B., O'Connor, N. and Murphy, N. 2004. Dialogue Scene Detection in Movies using Low and Mid-Level Visual Features. Intl. Workshop on Image, Video, and Audio Retrieval and Mining.

10. Smeaton, A.F, Gurrin, C., Lee, H., Mc Donald, K., Murphy, N., O'Connor, N., O'Sullivan, D., Smyth, B. and Wilson, D. The Físchlár-News-Stories System: personalized access to an archive of TV news. Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval (RIAO 2004).

11. Cooke, E., Ferguson, P., Gaughan, G., Gurrin, C., Jones, G., Le Borgne, H., Lee, H., Marlow, S., Mc Donald, K., McHugh, M., Murphy, N., O'Connor, N., O'Hare, N., Rothwell, S., Smeaton, A.F. and Wilkins, P. 2004. TRECVID 2004 Experiments in Dublin City University. Text REtrieval Conference TRECVID Workshop 2004.