

CANDELA – INTEGRATED STORAGE, ANALYSIS AND DISTRIBUTION OF VIDEO CONTENT FOR INTELLIGENT INFORMATION SYSTEMS*

P. MERKUS¹, X. DESURMONT², E.G.T JASPERS¹, R.G.J. WIJNHOFEN¹,
O. CAIGNART³, J-F DELAIGLE², AND W. FAVOREEL⁴

¹*Bosch Security Systems
Eindhoven, The Netherlands*

²*Multitel[†]
Mons, Belgium*

³*IT-Optics
Mons, Belgium*

⁴*Traficon
Bissegem, Belgium*

The introduction of digital video has led to a wide range of new video applications, including storage for information systems. Even though interactivity enables browsing and instant playback for such systems, the high information density and the large amounts of data result in cumbersome searching to find the information of interest. To solve this problem, the CANDELA project explores the combination of video content analysis, storage and retrieval for distributed systems. The concept of generating high-level content descriptions spans a wide range of new application. In this paper we elaborate on some parts by using the surveillance application as a pilot.

1. Introduction

Although many different types of technologies for information systems have evolved over the last decades (such as databases, video systems, the Internet and mobile telecommunication), the integration of these technologies is just in its infancy and has the potential to introduce “intelligent” systems. The CANDELA project, which is part of the European ITEA program, focuses on the integration of video content analysis in combination with networked delivery and storage technologies. To unleash the full potential of such integration, video-content analysis techniques are being explored. New algorithms for computer vision, innovative databases and high-bandwidth networks are being addressed.

2. Market relevance

After the introduction of DVB (Digital Video Broadcasting), video enhancement and interactive video added the possibilities for the user (the watcher) to interact with the video delivery system. Still the video signal itself is not “understood” by the system. The step towards intelligent video is the

* This work is part of the European R&D program in software-intensive systems, ITEA.

† This work has been supported by the Walloon Region.

addition of notion of the content by analysis of the video stream itself. The system then “understands” the video signal being delivered. This implies video content analysis techniques such as computer vision algorithms like segmentation into video objects, metadata generation for large databases of video content and the use of search and presentation devices as well as security mechanisms.

Currently, the development of digital video is mainly focused on state-of-the-art video encoding techniques (MPEG-4/H.264), describing the video content (MPEG-7), and standardizing a framework to enable interoperability (MPEG-21). All these standards are very much related to the scope of CANDELA, but do not address the application-specific requirements. For example, how can we detect a pulmonary embolism in the huge amount of pictorial data from a medical CT scanner? How can we detect and identify a shoplifter in a warehouse without manually observing hundreds of security cameras? How can we retrieve information about our favoured holiday place on a mobile device by applying abstract search queries on huge databases? How do we search through thousands of hours of video content from an airport in which a person abandons an object at a specific location?

The MPEG standards, in combination with the abundant availability of inexpensive digital processing power and large network bandwidth, enables analysis and delivery of video content in real-time at affordable cost. Hence, it offers unprecedented possibilities for application development in the areas of security, consumer entertainment, education and the like.

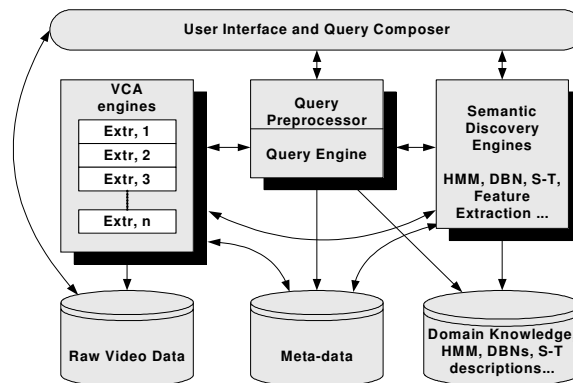


Figure 1. Example CANDELA applications, with content analysis, storage and delivery.

Figure 1 shows the general system architecture, which was already identified by Petrovic and Jonker [8]. It comprises the integration of content analysis, storage, querying and searching for many application domains like surveillance, medical imaging and home video. The instantiation of the Video Content Analysis (VCA), the databases (DBs) is domain specific though.

3. Example surveillance application system

In the previous sections we have shown the relevance of integrating existing technologies for a range of applications. On a system-level, these applications show a data flow to an equivalent set of subsystems. Let us now discuss the system architecture, adopting a surveillance system as a pilot application. Figure 2 schematically shows the system architecture of such application.

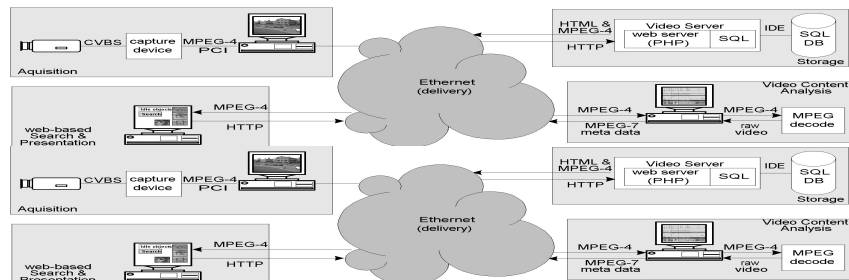


Figure 2. System architecture of a distributed surveillance application.

On the top-left side there is an observation environment consisting of security cameras, connected to a WAN (like the Internet). The DB, which is indicated in the top-right of the figure stores the video content. Note that this database could be distributed over the network. For example, the video content from an ATM machine in a bank is stored locally, but can be accessed from any other branch office. Dependent on the application, VCA has been applied in real-time to generate metadata that describes the content of the video. More specifically, this could be a rectangle to indicate the location and size of each moving object, an object classification to distinguish people, cars, etc., or even a persons name using face recognition and identification. The metadata from VCA is formatted and stored in the DB along with the corresponding video. Consequently, search queries on a high abstract level can be applied, using e.g. a web-based user interface on any remote client. For example, by drawing a line over a road in the picture, all video content is searched, in which vehicles cross that line. The following elaborates on the VCA, the metadata, and its data format.

3.1. Video content analysis

High-level interpretation of events within the scene requires low-level VCA of the image and of the moving objects. For our system, the architecture of the VCA part is divided in three main levels of computation to achieve the interpretation (Figure 3). The image level: acquisition; image filtering [1]; background evaluation; and segmentation. The blob level: description; blobs

filtering; matching; and tracking. The event level: tracking analysis and finite state machine. In the following we will only focus on the segmentation, description, tracking, and analysis.

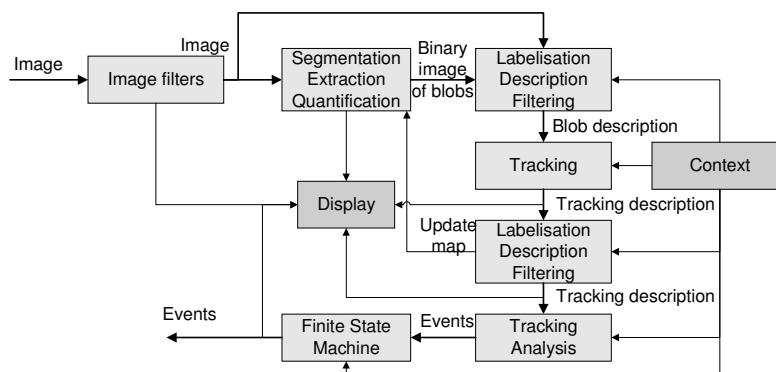


Figure 3. Block diagram of the video content analysis functionality.

3.1.1 Segmentation

The common bottom-up approach for segmenting moving objects uses both background estimation and foreground extraction [2]. The typical problems of this approach are changes of illumination, phantom objects, shadows of objects, camera vibrations, and moving tree branches. For representing the backgrounds we implemented several techniques such as low pass temporal recursive filter, median filter, mixture of Gaussians [3] and vector quantisation [4].

3.1.2 Blobs description and filtering

The aim of blobs description and filtering is to make the interface between foreground extraction and tracking and to simplify the information. The description process translates video data into a symbolic representation to reduce the amount of information that is necessary for the tracking module. The description process calculates, from the image and the segmentation results at time t , the k different observed features of a blob, e.g. 3D position in the scene, bounding box (rectangle around the object), mean RGB colour, blob shape, etc. To reduce noise artefacts, filtering is applied by removing small blobs, blobs outside the region of interest, etc.

3.1.3 Tracking algorithm

The tracking algorithm is divided in four steps that follow a straightforward approach: prediction, cost matrix computation, matching decision and tracks update. Note that there are multiple decisions when using MHT (multiple hypothesis tracking [5]). After the tracking process, the tracking description converts the internal tracking result to a graph, and adds some useful information to the matching data. It computes the time of life of every blob of a track. It also

annotates part of a track that is restricted to a small area, e.g. a stopped object. At the tracking description output, the tracking filtering is performed. To remove temporal inconsistencies due to noise, the detection data is filtered.

3.1.4 Tracking analysis and event generation

The tracking analysis is a process that receives the tracking description. It can find predefined patterns like objects entering from a defined zone of the image and exiting by another one, or objects that have exceeded a certain speed limit. Also objects that were stationary for a certain amount of time can be detected. Figure 4 shows this particular pattern of tracking description. In our application we are using this last pattern recognition as “someone is taking an object”. A similar approach has been described in [6].

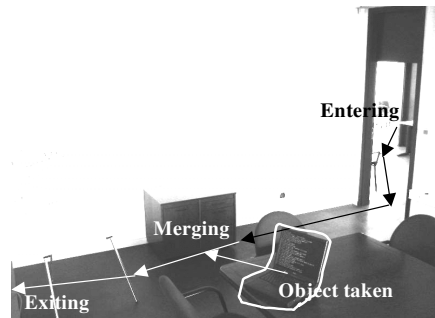


Figure 4. Tracking description pattern for “stolen object”.

3.2. Metadata analysis

The previous subsection has shown that video descriptions can automatically be generated using VCA algorithms. By exploiting this metadata, video content of interest can be found. However, typically, the descriptions from the analysis have a low abstraction level. For example, a moving object is described by a bounding box, and the track of the object. On a higher abstraction level, additional analysis is required to enable search queries in an intuitive way. This can be done by analysing the metadata generated by the above-described VCA algorithms. To obtain a higher semantic level of the metadata, a priori knowledge about the environment is required.

Petrovic and Willems [8] have already proposed a system that separates general VCA processing and additional knowledge for the environment that is application specific. To illustrate the additional analysis with a priori knowledge we elaborate on the following examples:

- perspective transformation to measure real-world sizes of the objects;
- object speed calculation;
- object classification.

To measure the real-world object sizes, perspective transformation needs to be applied since a standard security camera only provides data from two spatial dimensions, i.e. no depth information is available. To establish this, the algorithm has to be calibrated using additional information, e.g. the height of the camera, the position of two fixed points in the image and the angle of the ground floor. Figure 5 shows how the perspective transformation is used to calculate object sizes. Note that 'h' is the height of the camera, and 'x1' and 'x2' are the distances to two specific points in the scene.

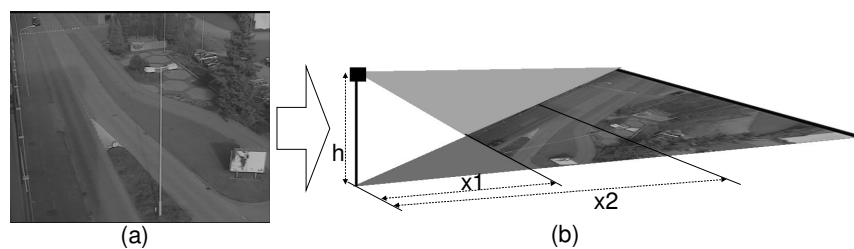


Figure 5 Captured camera image (a) and the calibrated camera image (b).

Having the perspective transformation available, the track of an object can be used to determine the speed. Hence the same a priori knowledge of the environment is exploited. Having the object sizes and speed, knowledge on the typical values of these metrics for objects like persons, cars, trucks and bicycles can be used to perform object classification.

On an even higher semantic level, analysis or reasoning can be applied. For example, a person that has been merged with a car that has speed zero implies that the person has stepped into the car. On the other hand, a person that has been merged with a car that has a speed that is higher than zero may imply an accident. Lets consider a second example. If a security guard has to check a building every half an hour, at least one person has to be detected on a repetitive basis within roughly the same time interval. The absence of person detection within the interval may imply that the guard has left his duty or was attacked by an intruder. Many contributions have been made to behavioural reasoning [9] [10].

3.3. Metadata format

As discussed in the previous subsections, VCA is applied to describe specific events in the video content. The descriptions extracted from the video data, i.e. the metadata, may be used to find requested video content from a search query.

CAVIAR [7] is the first de facto standard to store metadata. The main objective of CAVIAR is to address the scientific question: Can rich local image descriptions from video content improve image-based recognition processes. The main disadvantage of this de facto standard is the text-file format of the data structures. Although the produced metadata represent meaningful information or events, the format is not very suited for embedded systems.

Typically, a VCA algorithm provides descriptors from moving objects, such as their bounding box and their position at every frame. Besides these object features, the VCA may also signal events like “a person abandons an object”, including the time, place and actors of the event. Obviously, this type of metadata is very application specific. Therefore, it is important to adopt standardized formats to enable interoperability.

For the data structure of the metadata, XML is very suitable, since the type of metadata is arbitrary. Only the syntax is defined, whereas tags describe the semantics of the metadata, making it specifically suitable for application specific metadata. Although XML is standardized and very flexible, the broad applicability hampers the interoperability. As a result, ISO/ITU has standardized a rich set of tools to describe multimedia content using XML (MPEG-7). It describes the metadata elements, their structure and relationships to form the basis for efficient access to multimedia content. Both object features and events can be described easily. Typically, for an application such as video surveillance, only a subset of MPEG-7 is required.

Within CANDELA we have limited the set of the descriptors to satisfy our requirements. Figure 6 shows part of an output file of the above-mentioned “person abandons an object” application.

```
<xml>
<Event id='1'>
  <Texte>Object Left Behind</Texte>
  <TimeStamp>T00:01:14</TimeStamp>
  <TimeDuration>PT00:12:35</TimeDuration>
  <Position>
    <XCoord>0.92</XCoord>
    <YCoord>0.32</YCoord>
  </Position>
  <Bbox>
    <XExtent>0.12</XExtent>
    <YExtent>0.10</YExtent>
  </Bbox>
</Event>
<Event id='2'>
  ...
```

Figure 6. Example XML description for an abandoned object.

4. Conclusion

Although most parts of the system are based on state-of-the-art technology, the integration of video content analysis, storage, retrieval and delivery enables intelligent and powerful information systems. On a system level, the architectures are very similar, even though the range of applications can be very diverse. The example surveillance application shows that specific video content analysis algorithms are required to extract the application-specific metadata. To obtain a higher semantic level of metadata, a priori knowledge is added to the system. Besides bounding boxes and object tracks, also object type, size, and speed can be derived. To enable interoperability, MPEG-7 has been adopted as the format of the metadata. A demonstrator of the complete surveillance

application, including image acquisition, content analysis and retrieval in a distributed system, proving the feasibility, has been implemented and shown on the 5th ITEA Symposium in Seville, Spain.

References

- [1] J. Shen, S. Castan, "An Optimal Linear Operator for Step Edge Detection", CVGIP, vol.54, 112-133, (1992).
- [2] A. Elgammal, R. Duraiswami, D. Harwood and L.S. Davis, "Background and Foreground Modeling Using Nonparametric Kernel Density Estimation", Proceedings of the IEEE, vol.90, No. 7, July 2002.
- [3] C. Stauffer, W.E.L. Grimson, "Adaptive Background mixture models for real-time tracking", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pages 246--252, June 1999.
- [4] K. Kim, T. Horprasert, D. Harwood, L. Davis, "Codebook-based Background Subtraction and Performance Evaluation Methodology", 2003.
- [5] I.J. Cox and S.L. Hingorani, "An Efficient Implementation of Reid's Multiple Hypothesis Tracking Algorithm and Its Evaluation for the purpose of Visual Tracking", IEEE Transactions on pattern analysis and machine intelligence, Volume 18, Issue 2, February 1996, Pages: 138--150.
- [6] J.H. Piater, S. Richetto, and J. L. Crowley, "Event-based Activity Analysis in Live Video using a Generic Object Tracker", Projet Prima, Laboratoire GAVIR-IMAG, INRIA, Proceeding 3rd IEEE Int. Workshop on PETS, Copenhagen, June 1 2002.
- [7] <http://homepages.inf.ed.ac.uk/rbf/CAVIAR>
- [8] M. Petkovic, W. Jonker, Content-Based Video Retrieval, A Database Perspective, Series: Multimedia Systems and Applications, Vol. 25, 2003, ISBN:1-4020-7617-7.
- [9] M. Petkovic, W. Jonker, "An Overview of Data Models and Query Languages for Content-based Video Retrieval", International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet, L'Aquila, Italy, July 2000.
- [10] S. Hongeng, R. Nevatia, F. Bremond, "Video-based event recognition activity representation and probabilistic recognition methods", Computer Vision and Image Understanding, Volume 96, Issue 2, November 2004, Pages 129--162.