MULTIMEDIA KNOWLEDGE-BASED CONTENT ANALYSIS OVER DISTRIBUTED ARCHITECTURE

C. Carincotte[1], F. Bremond[2], J.-M. Odobez[3], L. Patino[2], B. Ravera[4] and X. Desurmont[1]

[1]Multitel, Mons, Belgium; [2]INRIA, Sophia Antipolis, France; [3]IDIAP, Martigny, Switzerland; [4]Thales Communications, Colombes, France

E-mail: [1]{carincotte,desurmont}@multitel.be, [3]odobez@idiap.ch, [2]{francois.bremond,jose-luis.patino_vilchis}@sophia.inria.fr, [4]bertrand.ravera@fr.thalesgroup.com

Abstract: **In this paper, we review the recently finished CARETAKER project outcomes from a system point of view. The IST FP6-027231 CARETAKER project aimed at studying, developing and assessing multimedia knowledge-based content analysis, knowledge extraction components, and metadata management sub-systems in the context of automated situation awareness and decision support. More precisely, CARETAKER focused on the extraction of a structured knowledge from large multimedia collections recorded over surveillance networks of camera and microphones deployed in real sites. Indeed, the produced audio-visual streams, in addition to security and safety issues, represent a useful source of information when stored and automatically analysed, for instance in urban planning or resource optimisation. In this paper, we overview the communication architecture developed for the project, and detail the different innovative content analysis components developed within the test-beds. We also highlight the different technical concerns encountered for each individual brick, which are common issues in distributed media applications.**

Keywords: communication architecture, applications for sensor networks, multimodal applications

1 INTRODUCTION

Advances in sensor devices, communications and storage capacities make it increasingly easy to collect large corpora of multimedia material. However, the value of this recorded data is only unlocked by technologies that can effectively exploit the knowledge it contains. It was thus the goal of the CARETAKER project [1] to investigate techniques allowing the automatic extraction of relevant semantic metadata from raw multimedia, to explore the value of the extracted information to relevant users, and to demonstrate this in real-scale configurations.

CARETAKER focused on the extraction of a structured knowledge from multimedia collections recorded over a network of camera and microphones. The motivation is that, despite the legitimacy of a number of privacy issues, such networks are becoming more and more common in different environments such as public transportation premises, cities, public buildings or commercial establishments, and that the multimedia streams of information they produce, in addition to surveillance and safety issues, could potentially represent a useful source of information if stored and automatically analyzed, for instance in urban planning and resource optimization applications. Indeed, the way the produced knowledge is used and shared in closed circuit television (CCTV) systems (and more generally in distributed multimedia applications) calls now for more considerations so as to bridge the gap between specific analysis algorithms and end-users expectations. For instance, a security operator may not only want to be informed in case of important event detection, but also to be able to rapidly and dynamically analyze the produced metadata to understand how/why an unexpected event has been identified. He may also want to analyze the generated metadata over a long period, so as to discover general pattern of events/activities within the monitored architecture.

Thus, the overall goal of the project was to investigate current and novel technologies to extract and exploit this information, and to evaluate them in a real test case, while exploring the added-value of this technology for real users. In this paper, we propose to present the audio/video distributed architecture developed within the project, and to highlight for each key component/application the innovative aspect of the conducted studies.

The reminder of this paper is organized as follows. A presentation of the global CARETAKER context is first achieved in Sec. 2. Sec. 3 then focuses on the data-streaming part of the architecture, from data acquisition, through server/client parties to the raw data storage aspect. . Sec. 4 and 5 then respectively detail the on-line and off-line metadata management and analysis subsystems. Sec. 6 then introduces the graphical user interface and the features implemented for the project purposes. Conclusions and perspectives are last drawn in Sec. 7.

2 ARCHITECTURE OVERVIEW

So as to detail the developed system, it is first necessary to delineate the functionalities it intends to take on. The CARETAKER monitoring environment aims at assessing multimedia knowledge-based content extraction and

Corresponding author: C. Carincotte, Multitel asbl, Rue Pierre & Marie Curie, 2 - 7000 Mons – Belgium, carincotte@multitel.be

analysis components in surveillance context, by focusing on both:

the reliable extraction of structured knowledge from data acquired over real-scale networks of camera and microphones (surveillance network of Roma/Turin metros). This was achieved thanks to the investigation of distributed techniques for ***real-time*** *extraction of semantic metadata* from audio/video raw data streams (e.g. [2, 7]);

the relevant exploitation of the extracted information to ease end-user missions (metro monitoring by safety/security operators), which was addressed through studies dedicated to ***off-line*** *processing of metadata* for extraction of long term patterns of activity (e.g. [3,4,8,9,10]).

Figure 1 illustrates the overall architecture of the CARETAKER monitoring environment in which both on-line and off-line subsystems (respectively the EVENT RECOGNITION SUBSYTEM and the KNOWLEDGE DISCOVERY SUBSYTEM) are identifiable.

2.1 Data acquisition and encoding

First, an acquisition subsystem is responsible for the audio/video data acquisition from the CCTV network (MPEG4-part2 video streams and raw audio streams). This brick is also in charge of the storage of the acquired data in dedicated database (SOLIDTech SOLID database engine), to allow post analysis and playback. Last, this component handles the delivery of live audio/video streams over the network using the standard Real Time Transport Protocol (RTP - implementation of the rfc 1889).

2.2 Event recognition subsystem

As depicted in Figure 1, the acquired audio-visual streams are streamed over the network and analyzed by different real-time modules.

In more details, raw data coming from the sensors are first encoded, stored and streamed over the network by the acquisition system (DATA ACQUISITION AND ENCODING module). Streamed data are then analyzed by a first processing unit responsible for the low-level features extraction (REAL-TIME LOW-LEVEL ANALYSIS MODULE). This layer allows the extraction of some primitive characteristics from the audio/video raw data such as ambient sounds, mobile objects, object trajectories... The low-level semantic descriptors (metadata) resulting from this analysis are then incorporated into the knowledge management system (AGENT-BASED DATA WAREHOUSE). This knowledge management system [5,6], which roughly corresponds to the database used for the storage of the metadata, is called Data Warehouse (DW).

A second layer of higher-level analysis (REAL-TIME HIGH-LEVEL ANALYSIS MODULE) then processes the previously computed metadata, in conjunction with the audio/video streams, so as to identify events of interest, such as turnstile jumping, abandoned luggage detection... The resulting high-level metadata is also incorporated in the DW.

2.3 Knowledge discovery subsystem

Regarding the offline part of the architecture, a *knowledge discovery* module (OFF-LINE KNOWLEDGE DISCOVERY MODULE) analyses the stored metadata using clustering and data mining techniques. The aim of this component is to identify general trends in the stored metadata, computing statistics (flow of people, space usage...) and to explore the relationship between different types of events.

2.4 Subsystems graphical user interfaces

With respect to both on-line and off-line processing, the extracted information is exploited through two dedicated subsystems.

The first one, i.e. the EVENT RECOGNITION SUBSYSTEM, offers the standard monitoring interface triggering alarms corresponding to events detected in real-time. This subsystem also allows event driven retrieval of the audio/video data and corresponding metadata for inspection purpose.

The second one, i.e. the KNOWLEDGE DISCOVERY SUBSYSTEM, allows users to query combinations of higher-level semantic events, to run unsupervised clustering and data mining algorithms on the stored metadata, to compute statistics about space usage...

2.5 Metadata structure and exchange

As highlighted in Figure 1, the system has to handle three kinds of metadata: results from analysis (performed in low and high-level analysis modules), queries and replies (mediated by both on-line and off-line interfaces).

In order to guarantee system consistency and compliance with standards, these three kinds of metadata share the same markup language, i.e. the eXtensible Markup Language (XML). Furthermore, in order to avoid sending raw XML documents over the network, every transferred data is wrapped in an RSS feed. The metadata exchange is thus reduced to the handling of an RSS flow, which enables additional fields, such as producer identifiers, technical data...

3 DATA ACQUISITION & ENCODING

This section provide an overview of the audio/video acquisition module developed within the project, as well as the raw data storage and streaming architecture related to it.

Data encoding The CARETAKER system has been built up to be deployed in two different test sites. In order to

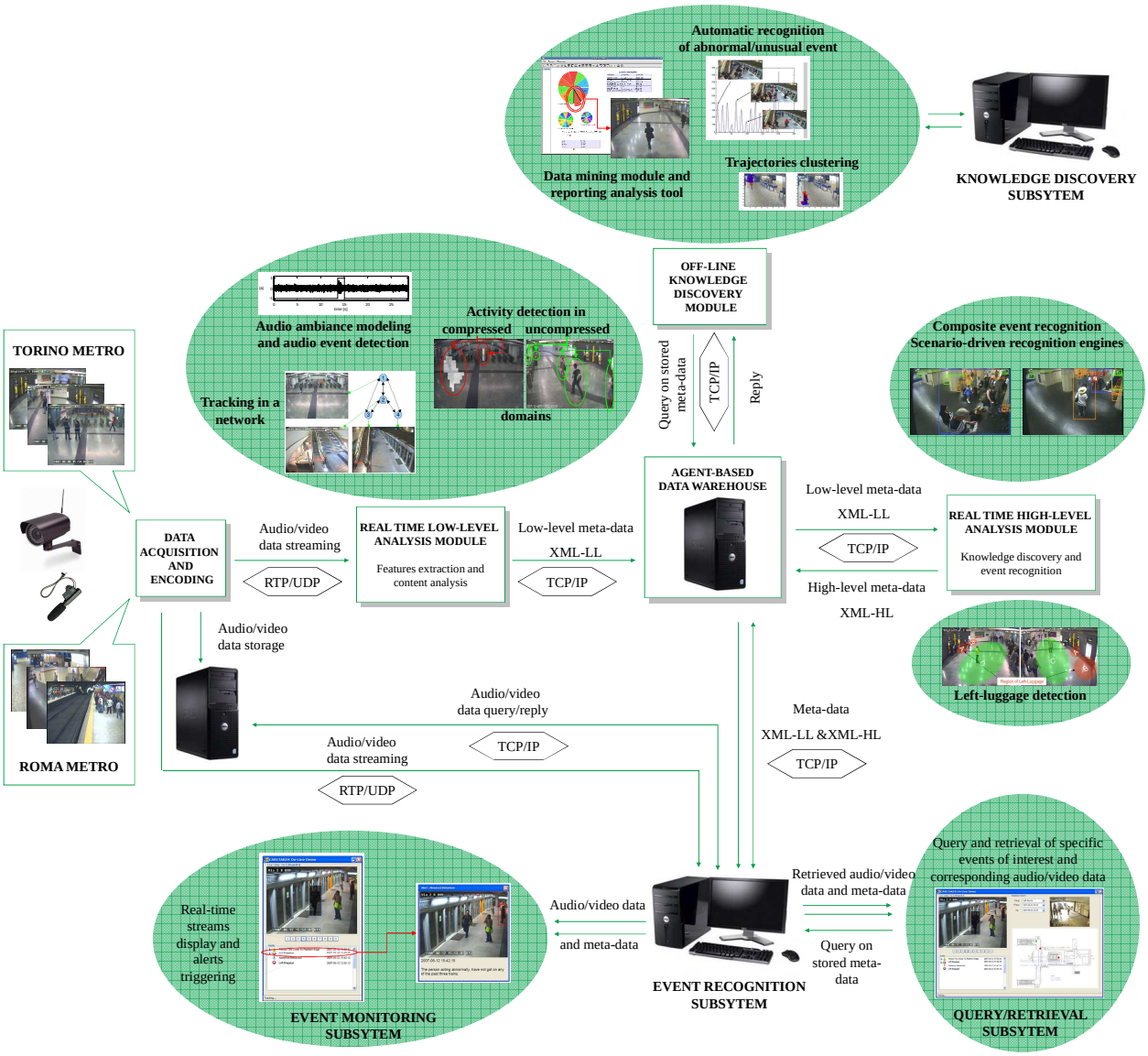

Figure 1 Design configuration of the CARETAKER monitoring system.

provide a consistent system, a common acquisition framework has been developed to handle the two sites specificities (Roma acquisition platform is made of a IEI 8371-P MPEG4 video acquisition board for the video part, and an external EDIROL UA25 USB audio acquisition board - Turin architecture is based on the pre-installed video streaming system provided by the CCTV network manufacturer of Turin underground).

Dynamic network management One of the main issues related to already deployed CCTV system is compliancy with audio-video standard compression (MPEG4, H264, Audio PCM, …) and also with IT network. After digitalisation, audio/video acquired streams are identified thanks to a set of IP addresses. This technical key point allows interfacing the system with various digital networks bypassing the digitalisation system. The only open issue is the dynamic IP address management. A specific tool has been developed in order to maintain a dynamic index of the entire sensors. Thank to this dynamic tool, CARETAKER can be easily interfaced with various deployed systems (analogue and digital). Due to time constraints, the current system accepts only MPEG4 an Audio PCM format as digital inputs. The main issue is here to interface the system to full compliant standard streams. In the case of non standard audio video streams, adapted decoders have to be integrated. This would remain currently the only difficulty to interface the CARETAKER system.

Data synchronization To enable a reliable synchronization between live data and stored data, each live audio/video frame is time stamped using a 64 bit integer (the time format is based on the ISO8601 standard). Each recorded audio/video data is also time-stamped when it is written to the database.

Data storage In order to write the encoded video data and raw audio data to the solidDB database, the system splits them into manageable chunks during the acquisition. Insertion of the chunks is then performed via the database client API. The database schema is thus very straight-forward, consisting of 5 principal tables. The technical characteristics of the video part

of the stream (frame rate, codec, image height and width) are stored in the `video_header` table. Likewise, the audio technical characteristics (sample rate, codec and bits per sample) are stored in the `audio_header` table.

Sensor ID specification Regarding the sensor id, stored streams are identified as coming from a single audio/video device. Prior to retrieve and playback streams from the database, the GUI/off-line tools have thus to parse a lookup table, responsible for mapping the physical sensor id (which are the real security system sensors used by processing) and the streams id used for the storage.

Data streaming To allow content analysis bricks to receive MPEG4 and raw audio data, a generic software API has been developed. In this way, the online analysis modules (responsible for the live streams processing) are able to connect to the acquisition system using an RTP/UDP link. As the UDP transport protocol allows the use of IP multicast option, the server is able to deliver the live video and audio stream to an unlimited number of clients. The only requirement for clients is that they must be a member of the specified "multicast group".

4 EVENT RECOGNITION SUBSYTEM

Some very innovative approaches have been developed within the project to provide a useful understanding of the infrastructure activities and respond to user needs.

Several of these activities were pertaining to the audio domain (recognition of basic audio events -e.g. train arrival, announcement- or abnormal ones -unusual shouts), but most of them were based on video analyses. Amongst others, one can cite: activity modeling (2D object recognition, e.g. unauthorized dogs [14] or bicycles [13]; ticket-vending machine queue detection [4], left-luggage detection, train-stop detection measures, and platform occupancy levels measurements [3], jumping over turnstiles), using either statistical or ontology driven approaches [7]; single or multiple person tracking, using single or multiple cameras, e.g. to allow a end-user to tag a person and track him automatically in the infrastructure [2,15]; finally different knowledge discovery algorithms have been implemented for building statistics about the infrastructure usage (See Sec. 5). Most of the algorithms were tested on both metro sites.

Overall, the proposed communication system proved to be sufficiently flexible to implement the above algorithms, as many of them worked on single streams. However, some of the tools needed specific considerations. As an important example, tracking with multiple cameras, whether with or without field-of-view overlap, required the implementation of circular buffers to ensure processing of the synchronized video sequences.

5 KNOWLEDGE DISCOVERY SUBSYSTEM

The Knowledge Discovery (KD) subsystem takes its input from the Data Warehouse (DW). Specific queries are automatically built from the system employing XML language. These are wrapped in a RSS feed and sent to the Data Warehouse. The KD system queries the metadata related to detected mobile objects and video and audio events occurring inside an observation period specified by the end-user. The answer to the query comes to the KD system as well on the form of an RSS feed where the information is made explicit in a frame by frame basis. In order to have a clear and compact representation of the human activity evolving on the scene, and with the aim to achieve knowledge discovery, the metadata is structured and saved into a dedicated KD database (which also employs SOLIDTech SOLID database engine) in the form of two different semantic tables: mobile objects table, events table. Apart for reordering the information in agreement with our semantic representation, there are a series of new fields calculated in order to extract new information. Off-line we calculate, for instance, the shape, the significant event involving a detected mobile.

Next, two clustering processes are applied to derive the knowledge from the streams of data. First, agglomerative hierarchical clustering [11] is used to characterize motion from mobile objects and extract the main flows of people and space occupancy in the underground. Secondly, relational analysis clustering [12] is employed to extract the relationship between people, and occurring video and audio events. For instance, we have detected the main flows of people when turnstyles are busy. Knowledge discovery results are again stored via SOLIDTech routines into the KD database.

6 GRAPHICAL USER INTERFACES

This section presents the requirements, design and implementation of the graphical user interfaces. First, the on-line interface, which allows end-users to interact with the on-line analysis tools, is presented.

Requirements and scope The on-line tools are characterized as those which provide the end-user with timely information about events or activities that have recently happened or are still on-going. The motivation for the delivery of this information is that the nature of these events or activities is such that the end-user may need to take immediate action in response. Thus, the use-case scenario is one in which the On-Line Graphical User Interface (OLGUI) is available to the operator alongside the standard surveillance streams and controls for camera selection and steering.

The main objectives for this interface are two-fold: firstly it must provide the necessary mechanism to start each of the analysis tools, on request from the user. Any information

required by the tool at this point must also be provided. Secondly, the User Interface must display the results of the analysis in a suitable manner.

A last important requirement was to make the GUI easily configured for different sites, e.g. Roma and Turin sites.

Design So as to allow the user to switch between the available streams in a straightforward and intuitive manner, a dynamic hierarchical menu structure was chosen. The structure of submenus and elements is provided by an xml configuration file. In addition to grouping the cameras by station, an intermediate sub-group of cameras was defined, to enable all cameras situated in a common physical area (e.g. a hallway or a platform) to be grouped together (see Figure 2).

Site-specific configuration The GUI is designed to work at different sites. Thus, configuration files are used to adapt the GUI to a specific arrangement of cameras and tools that are present. A first xml file stores the arrangement of available cameras (in groups and sub-groups) together with the associated calibration data for those cameras for which it is available. A second xml file stores the list of available analysis tools, together with the necessary associated data, i.e. the IP address of the host machine, port number for this analysis tool, and the list of sensors (cameras and microphones) which are valid input for this analysis tool. Thus, these elements of the GUI are configured to a specific surveillance site by means of these xml files, which while ensuring the GUI consistency between different sites, allow to adapt the GUI display to each site specificity. Figure 2 presents Roma GUI, in which the site-based areas are highlighted.

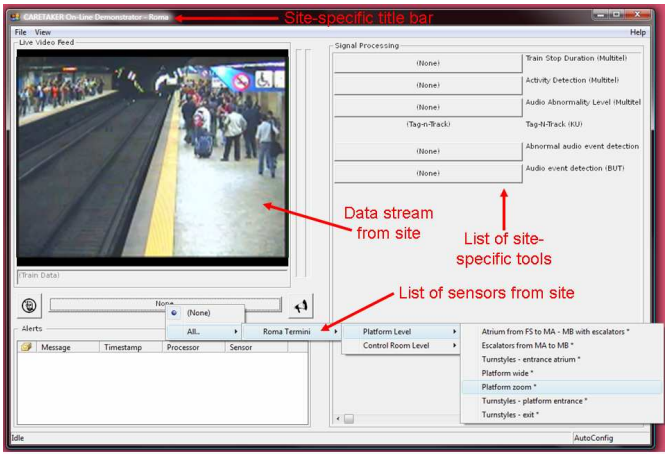

Figure 2 GUI site-based areas handle by configuration files.

For the event detection tools, in general they require as input the specific data stream that the tool is to be analyzing. This could be an audio or video stream, depending on the type of tool. In addition further constraints could be in place for particular tools that are configured only to use a subset of the available streams (e.g. turnstile jumping detection only available for turnstile views, train-stop monitoring only available for platform view). One design choice was to use the currently viewed stream as the implicit choice for the input to the processing tool, but the ambiguity over audio and video (and also over tool-specific constraints) suggested that it would be preferable to use a dedicated stream selection control for this purpose. Therefore, the same type of control (as used in the display stream selection) is used in the analysis stream selection (see Figure 3). Some specific developed tools, such as tracking a person from multiple cameras, have an additional window display (see Figure 4).

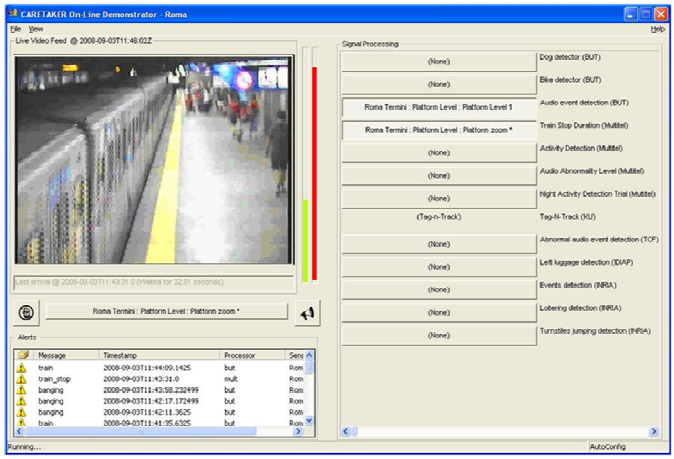

Figure 3 GUI event detection tool.

Figure 4 GUI tag and track tool.

Off-line interface We have developed a graphical tool where the end users select a period of recording time, which they want to interpret. A security operator may not only want to be informed in case of important event detection, but also to be able to rapidly and interactively analyze the generated metadata over a long period, so as to discover general pattern of events/activities within the monitored architecture. Figure 5 presents the off-line interface for people trajectory characterization; while Figure 6 presents the off-line tool giving the correlation between events and their statistics.

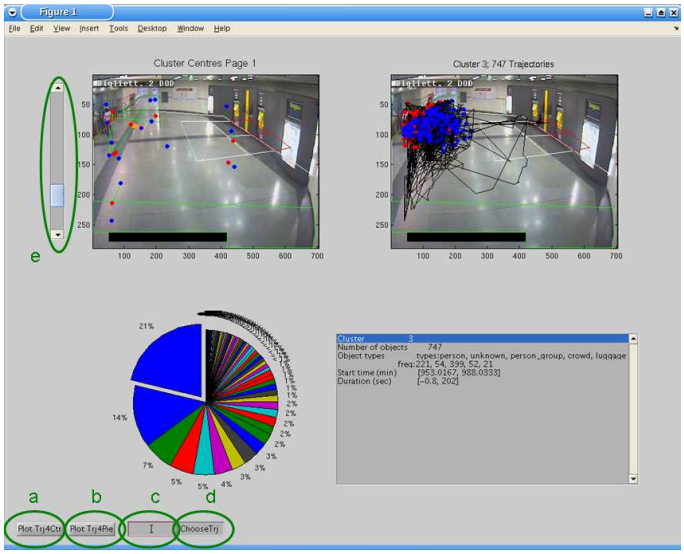

Figure 5 Off-line interface for trajectory. Buttons a to e allow to encode Solidsql (SOLIDTech) queries to explore different levels of flows of people and different station areas/equipements.

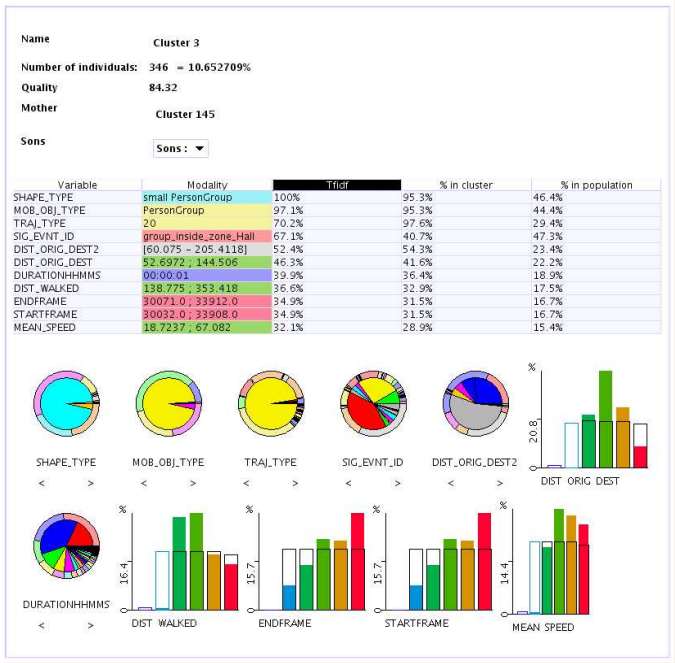

Figure 6 Off-line interface for events; end-users can analyse different activities with their statistics though pie-chart or bar-plot.

7 CONCLUSION

The CARETAKER project has successfully developed new multimedia knowledge-based content analysis techniques for knowledge extraction components, and metadata management sub-systems. More precisely, CARETAKER innovated by establishing algorithms for reliable extraction of structured knowledge from live streams of video and audio data acquired over the entire networks of two underground-transport sites: Rome and Turin. End-user needs and expectations were further satisfied by including a knowledge discovery subsystem extraction of long term statistics and patterns of activity, which can help on usage interpretation and management/optimization of the station. Appropriate storage and communication processes have been implemented for an efficient system running.

Acknowledgment

The work presented here is partially supported by the European Commission under the 6th Framework Program through the IST FP6-027231 CARETAKER project. For further information about the CARETAKER project, please visit http://sceptre.king.ac.uk/caretaker/.

References

[1] IST FP6-027231, CARETAKER: Content Analysis and REtrieval Technologies to Apply Knowledge Extraction to massive Recording, 2006-2008, http://sceptre.king.ac.uk/caretaker/.

[2] Jian Yao and Jean-Marc Odobez, Multi-Camera Multi-Person 3D Space Tracking with MCMC in Surveillance Scenarios, in European Conference on Computer Vision, workshop on Multi Camera and Multi-modal Sensor Fusion Algorithms and Applications (ECCV-M2SFA2), Marseille, Oct. 2008.

[3] C. Carincotte, X. Naturel, M. Hick, J.-M. Odobez, J. Yao, A. Bastide and B. Corbucci. Understanding metro station usage using Closed Circuit TeleVision cameras analysis. 11th Int. IEEE Conf. on Intelligent Transportation Systems (ITSC): p. 420-427. Beijing, China. October 2008.

[4] X. Naturel and J.-M. Odobez, Detecting Queues at Vending Machines: a Statistical Layered Approach, in IEEE Proc. Int. Conf. on Pattern Recognition (ICPR), Tampa, December 2008.

[5] C. Carincotte, X. Desurmont and A. Bastide. Adaptive metadata management system for distributed video content analysis. Advanced Concepts for Intelligent Vision Systems (ACIVS): p. 334-345. Juan-les-Pins, France. October 2008.

[6] B. Lienard, A. Hubaux, C. Carincotte, X. Desurmont and B. Barrie. On the Use of Real-Time Agents in Distributed Video Analysis Systems. Real-Time Image Processing, part of the IS&T SPIE Symposium on Electronic Imaging. San Jose, CA USA. January 2007.

[7] F. Cupillard, F. Brémond, M. Thonnat, Automatic Visual Recognition for Metro Surveillance. Proc. Int. Conf. Measuring Behavior, Wageningen The Netherlands, 2005

[8] L. Patino, H. Benhadda, E. Corvee, F. Bremond, M. Thonnat. Extraction of Activity Patterns on large Video Recordings. IET Computer Vision, Volume 2, Number 2, page 108--128 - june 2008.

[9] L. Patino, E. Corvee, F. Bremond, M. Thonnat. Data mining for activity extraction in video data. EGC 2008, Sophia Antipolis, France, page 433--444 - 29th January -1st February 2008.

[10] L. Patino, E. Corvee, F. Bremond, M. Thonnat. Management of large video recordings. 2nd International Conference on Ambient Intelligence Developments, AmI.d 2007, Sophia Antipolis, France, page 79--91 - 17th - 19th September 2007.

[11] L. Kaufman, J.P. Rousseeuw. Finding groups in data, Wiley-Interscience, 1990.

[12] F. Marcotorchino, P. Michaud. Optimisation en analyse ordinale des données, Masson, 1978.

[13] V. Beran, A. Herout, I., Řezníček: Video-Based Bicycle Detection in Underground Scenarios, In: Proceedings of WSCG'09, Plzeň, CZ, 2009, p. 4

[14] R. Juránek, Detection of Dogs in Video Using Statistical Classifiers, In: Proceedings of International Conference on Computer Vision and Graphics 2008, Heidelberg, DE, Springer, 2008, p. 11.

[15] A. Colombo, J. Orwell, S.A. Velastin, "Colour Constancy Techniques for Re-Recognition of Pedestrians from Multiple Surveillance Cameras", ECCV workshop on Multi-camera and Multi-modal Sensor Fusion Alforithms and Applications (M2SFA2), Marseille, France, October 2008.