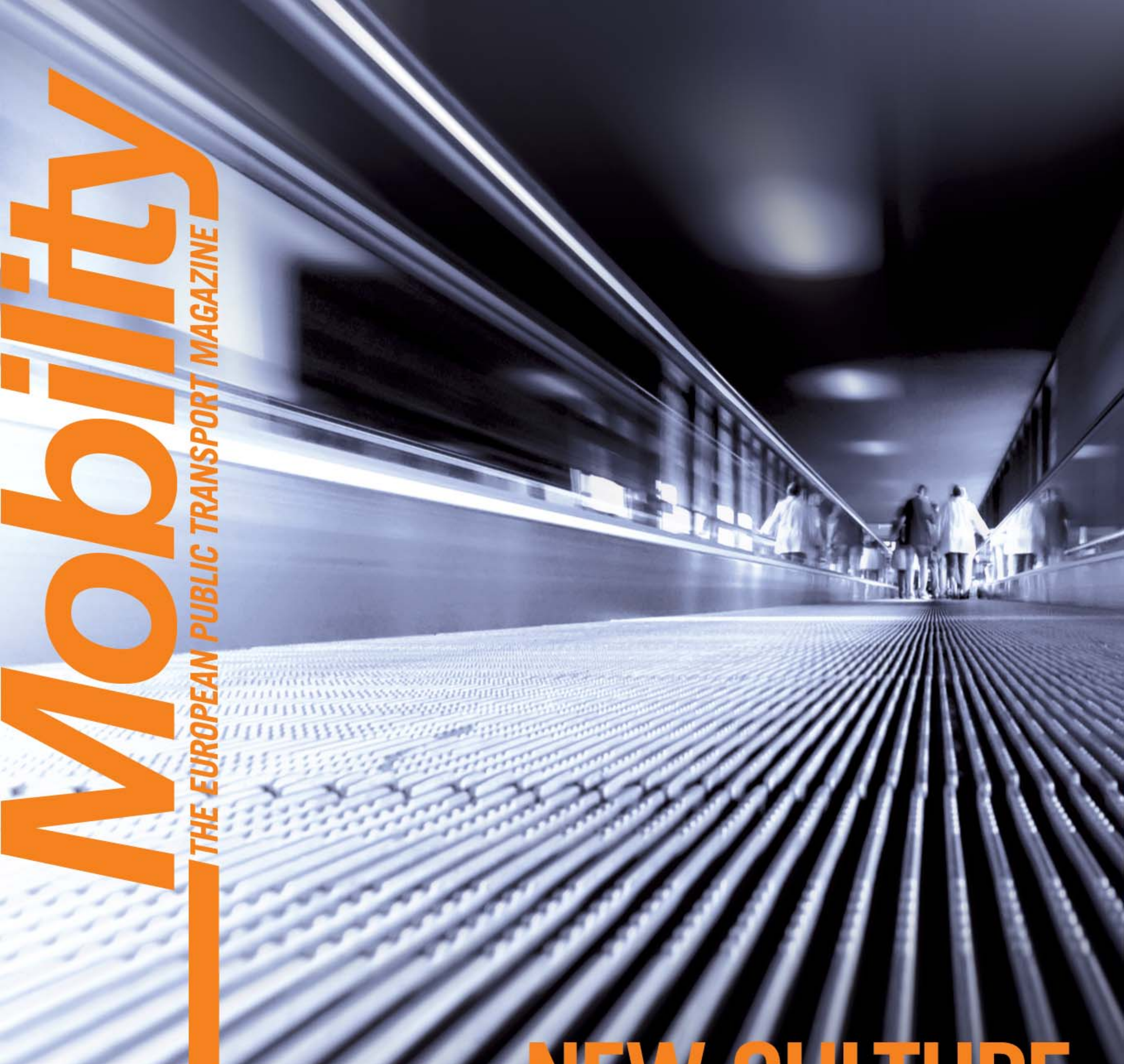
Mobility
THE EUROPEAN PUBLIC TRANSPORT MAGAZINE



NEW CULTURE FOR URBAN MOBILITY

ITS
Putting technology to the test

TICKETING
Smartcards gain currency in Europe

BUS
Taking the design route

ISSUE 13 www.mobility-mag.com
EUROPE: €26 - REST OF THE WORLD: €32

ITS

CARETAKER PUTS KNOWLEGE TO GOOD USE

Security is a fairly recent concern for transport companies. Urban transport used to be relatively secure and no-one could have imagined such an increase in crime and threats on board in the recent past. Closed Circuit Television (CCTV) is now an everyday part of our lives, at least in Europe. It helps to improve various daily operations and tasks, as well as providing users with various degrees of security.

It is not surprising, therefore, that a newborn metro such as the one in Turin installed more than 800 cameras for less than 10km of line and 15 stations, or that the vast and crowded London Underground has more 6,000 cameras.

But the question now is: who takes care of all these videos?

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 is thus the goal of CARETAKER, a European-funded research project, to investigate techniques allowing the automatic extraction of relevant semantic metadata from raw multimedia, explore the value of the extracted information to relevant users and demonstrate this in a framework that preserves the privacy of the individual.

More precisely, CARETAKER focuses on extracting structured knowledge from multimedia collections recorded over a network of camera and microphones installed in the undergrounds of Turin and Rome. The motivation is that, despite the legitimacy of a number of privacy issues, such networks are becoming increasingly common in different environments, e.g. public transport 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, e.g. in urban planning and resource optimization applications.

The project considers two types of content knowledge: a first layer of primitive events that can be extracted from the raw data streams, e.g. ambient sounds, the degree of crowding present in the scene, and the routes taken by individuals. A second layer of higher semantic events is defined after longer term analysis and is based on more complex relationships between the primitive events. It is important to note that while few real systems are equipped with such content extraction and analysis tools, academic laboratories have developed many algorithms partially addressing these issues, but most of the time applied on toy problems, with very little actual data or acted data. Thus, the overall goal of CARETAKER is to investigate current and novel technologies for extracting and exploiting this information, by evaluating them in a real test case, while exploring the added value of this technology for real users, starting with public transport operators.

Since CARETAKER is a multidisciplinary research project focused on public transport safety and security issues, it has given rise to major scientific and technological challenges (knowledge modelling for scene understanding, distributed surveillance systems for event recognition, data analysis for infrastructure and/or environment planning and resource optimization, etc).

CARETAKER is being carried out by a consortium of companies and research centres, from all over Europe, comprising two public transport operators – GTT, which runs Turin's automatic driverless metro and the Mobility Agency of Rome (ATAC), which manages the city metro through its subsidiary Met.Ro.); industrial partners such as Thales Communications and SOLID Information; technology and research centres and universities such as Multitel asbl, the Institut Nationale de Recherche en Informatique et en Automatique (INRIA), Kingston University (KU), Fondation de l'Institut Dalle Molle d'Intelligence Artificielle (IDIAP) and Brno University of Technology (BUT).

Multitel asbl, active in major European initiatives in scientific and technological domains, has a strong background in surveillance-like and monitoring applications. Its work focuses on developing and implementing innovative projects in collaboration with local and international companies, and promoting the acquired know-how through technology transfer to private companies fully related to Multitel. In the case of CARETAKER, the spin-off in question is ACIC s.a., which customizes software and hardware video analysis solutions for surveillance, dealing, for example, with intrusion detection, perimeter monitoring, abnormal behaviour or abandoned object detection, tailgating detection and unicity check, statistics gathering (counting, speed, classification), etc.

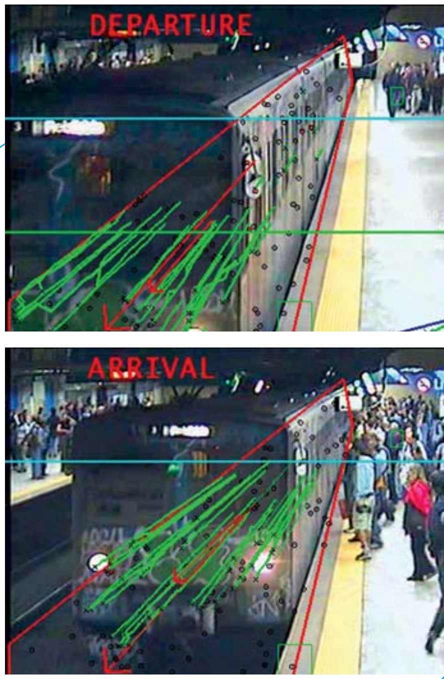
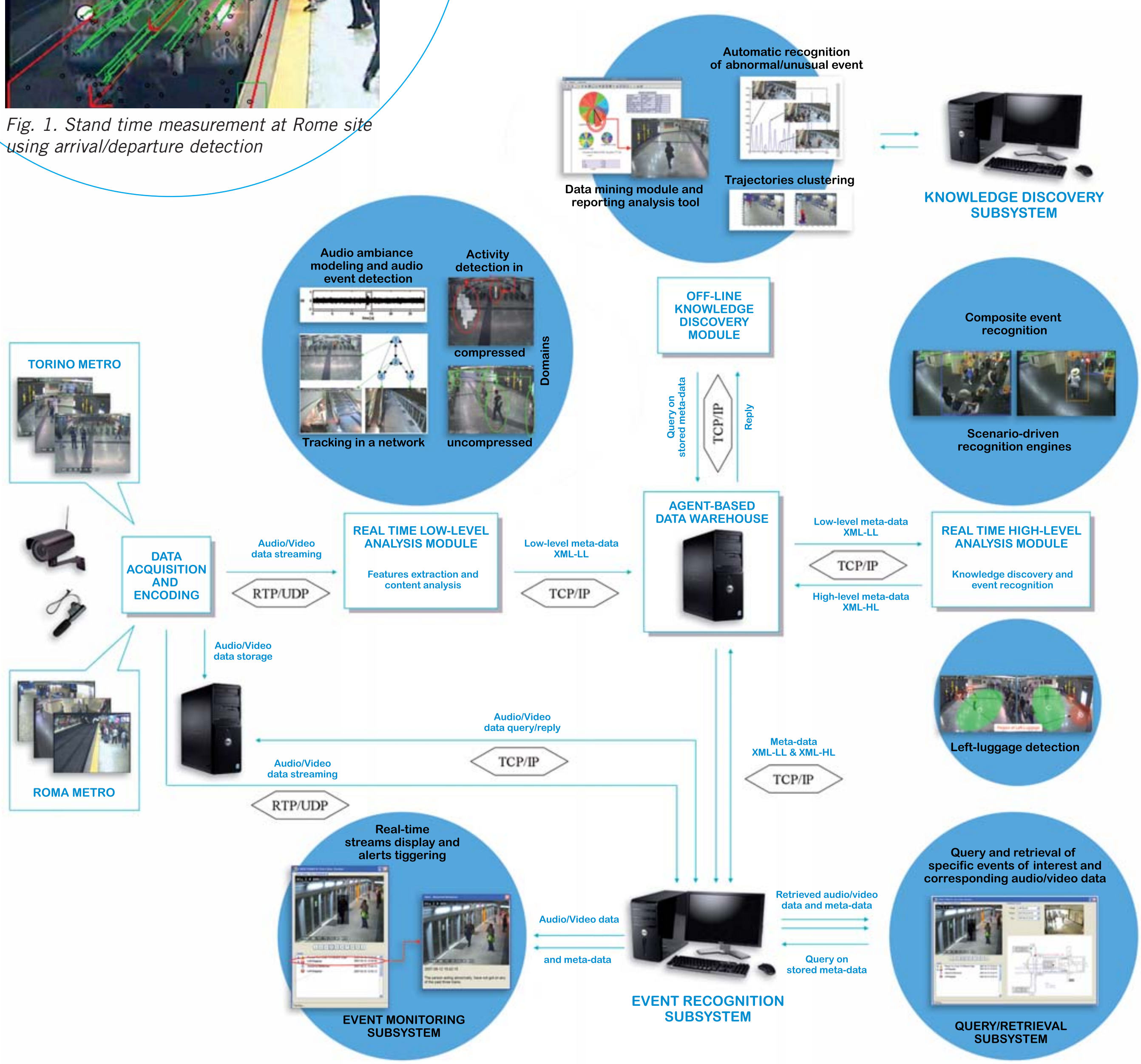


Fig. 1. Stand time measurement at Rome site using arrival/departure detection

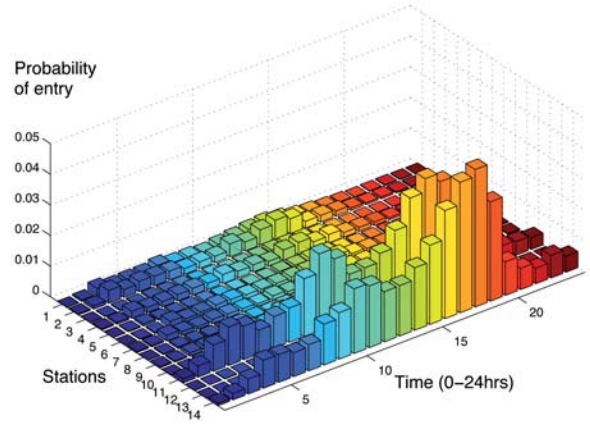
As such, Multitel is involved in the development of both audio/video analysis components and technical sub-systems. More precisely, it is currently working on the development of low level feature extraction modules (e.g. detection of moving people in video, audio stream analysis for abnormal situation detection…), and the integration of developed technologies into real-time demonstrators within the monitoring systems of Turin and Rome.

Thanks to its close collaboration with its spin-off, Mutitel is also assessing a statistics gathering tool provided by ACIC that enables, among other capacities, measurement of the stop-time of metros using the CCTV system already available in the Rome network (see Figure 1). This analysis is extremely revealing, e.g. it establishes a correlation between the number of passengers in the station (computed by another partner, IDIAP) and the related stop duration of metros. Another key impact of such analysis is that the added-value of the extracted information can then be used to obtain relevant useful statistics for infrastructure management, planning and optimization.

Another application useful for public transport operators and addressed by CARETAKER is the 'tag and track' application, discussed below. The motivation for the application is the need, by transport surveil-



lance system operators, to keep a particular individual 'in view' during the time they spend in the network. For example, this may be because they are previously known to the surveillance personnel, or because they are a vulnerable person travelling late at night. It is possible for the surveillance operator to manually switch cameras to keep the person in view, but this means that they are not able to attend to other tasks simultaneously. An automated system to follow this individual, which only requests operator attention when necessary, would save operator time and enable them to work more efficiently and effectively.



Relative frequency of passengers entering stations at different times of the day

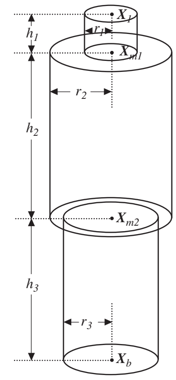


Fig. 2. 3D human representation

In a public transport network, such as the metro in Turin managed by GTT, the majority of passengers follow fairly predictable paths. Their route to and from platforms is well-defined and, for simple metro networks, offers little scope for variation. The mean journey times between stations, and the extent of any deviation from these times, are generally available. Statistically, passengers are more likely to have particular destinations at certain times of the day. The graph below shows statistics on the frequency of passengers entering each station at different times of the day, obtained using electronic data captured from gate entry points. Computer Vision techniques can be used to estimate the relative frequency at exit points, and also to build up statistics about the likely transitions between stations. The factors combine to create a relatively predictable environment for the movement of passengers in the network. This knowledge about people's movements constrains the "a priori" state uncertainty of the system. Any information derived from Computer Vision tracking applications will reduce this uncertainty still further.

To 'tag and track' successfully, two types of Computer Vision tools can be used. Firstly there are appearance-based descriptors, such as the MPEG-7 colour descriptors used to measure similarities in appearance between two observations of a passenger taken at different times and places. Secondly there is spatial tracking technology, to track the location of a passenger as they make their way between the ticket halls and platforms, and in and out of trains. Furthermore, the basis for the evaluation is being developed to predict how often the operator will need to guide the system, and therefore how much time the technology can save.

Tracking is an essential component of a scene understanding system. In CARETAKER, we address two main tasks related to this issue. The first is single person tracking over a network of cameras; the second is multiple object tracking (MOT), all using single or multiple cameras.

The single person tracking task directly answers some user needs, and has been coined the 'tag and track task'. More specifically, surveillance operators would like to be able to point at and 'tag' some target people or groups of people (e.g. a drunk person or people shouting and annoying other passengers) and track them through the metro station, so that there is always a monitor that displays a view of them. This type of awareness and camera selection mechanism is very important for control rooms, where the number of monitors is much less than the actual number of cameras in the network (e.g. 24 video monitors for more than 800 cameras in Turin metro). On the other hand, the MOT task is essential for the recognition of higher level events and monitoring of activities within the metro (e.g. how many people used the vending machine, who forgot their luggage, etc).

There are several existing techniques that tackle this task and/or tested on some simple surveillance sites. They are often 2D approaches based on identifying the 'foreground' regions of an image that do not correspond to the background scene, then attempting to decompose them into individual people segments. The last task is usually quite difficult to solve in practice, and this is why the foreground regions are usually classified into 'individual people', 'group of people', and 'crowd', according to location, size and shape.

In CARETAKER, IDIAP investigated alternative techniques based on 3D models. More precisely, the goal is to detect and track all individuals in the scene, with each individual represented in 3D using a set of elliptic cylinders (see Figure 2). And for each individual, we want to immediately estimate their 3D position on the ground, his speed, height and body orientation. Such a 3D approach clearly presents several advantages over 2D methods. Firstly, the parameters have a clear semantic interpretation (in 2D, the parameter for estimating would be: a position and size in the image plane). Secondly, using well formulated, probabilistic approaches, prior information on the tracks (e.g. people enter and exit from specific locations, people should not occupy the same 3D space, people move smoothly in the 3D space) can be incorporated into the model. And thirdly, using the camera calibration parameters, knowing the location of people in the scene (e.g. where they stand on the ground plane) allows to decide which person stands in front of whom when viewing the scene from a given camera (i.e. occlusion situation), which helps provide useful ways of modelling how should a video image look like given the position of people in the scene. Of course, solving the inverse problem (finding the best people localization parameters given the video measurements) is one of the hardest problems in computer vision, and specific optimisation techniques need to be exploited to address this issue. Nevertheless, in the project we have achieved significant progress with respect to the state-of-the-art.

Audio event detection

While video processing is routinely used in surveillance applications and already benefits from many years of research and technology development, the use of audio is still in its early stages. CARETAKER puts an emphasis on audio as we believe that its acquisition and processing can deliver increased quality and new features to automatic surveillance systems. Firstly, there are events such as gun shots or loud screams that can be detected only from audio. The second application is in joint processing of audio and video information, where the two modalities can enhance each other and provide complementary information.

The first issue comes with the acquisition of data. While there are many cameras in public places, in CARETAKER, it was necessary to install appropriate audio sensors at both sites (Rome and Turin) and negotiate with metro operators, install the cabling etc. For some of the applications studied, a hand annotation of part of the data was necessary (see Figure 3). Four hours of data from Roma were processed in this way, requiring an effort of several tens of hours of labor. The acquired data comes from:

1. Normal operation of the station.
2. Acted scenarios where sounds of interest were recorded (shots, screams, crashing into objects).

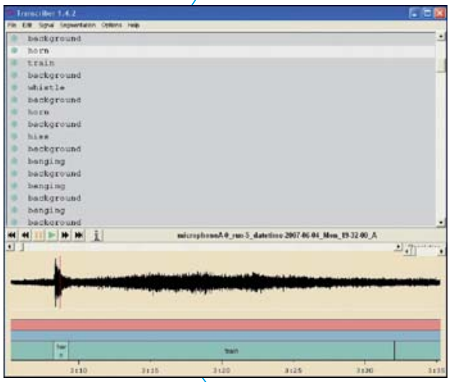


Fig. 3. Hand annotation of audio in CARETAKER

The problem in audio is that it does not allow for "Mosaicing": while it is possible to feed the operator with several video streams (separate monitors or windows) a mix of audio from several microphones is not audible. This calls even more for automated processing.

Processing in CARETAKER can be divided into two categories:

- automatic detection of known audio events such as normal audio ambiance, approaching train, closing doors, normal speech or announcements,
- detection of unusual events that were not previously seen and that are mostly the ones relevant for the surveillance scenario (loud scream, shot, strangely sounding train).

The audio processing in CARETAKER is investigated by three partners: Thales Communications [France] (the project coordinator), building on its extensive experience in speech coding and multimedia processing, Multitel [Belgium], which is largely experienced in audiovisual processing for surveillance and Brno University of Technology [Czech Republic], of which the Speech@FIT group is one of Europe's most advanced research groups in the area of speech processing.

The algorithms start with signal processing that aims at mimicking the human ear. Figure 4 shows an example of the Gammtone filter-bank that simulates human cochlea. Other tools include standard signal processing techniques such as Fast Fourier Transformation (FFT), modelling of long temporal trajectories and vector quantization. For classification and recognition of events, CARETAKER researchers have taken advantage of advanced tools such as artificial neural networks, hidden Markov models (HMM), tree-based clustering and techniques inspired by video processing such as morphological opening and closure.

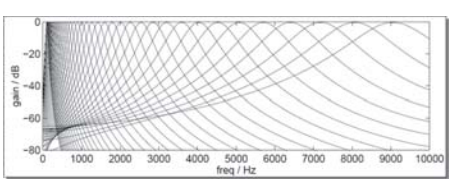


Fig. 4. Gammatone filter-bank processing audio

The audio applications investigated in CARETAKER are:

- Normalized Auditory Attention Levels for Automatic Audio Surveillance investigated by Multitel. The idea is that suspicious sounds are rare, so we need to suppress all usual sounds and emphasize the suspicious ones. Technically, this is done by detecting sudden and unexpected changes in audio textures with no "a priori" sound models using a purely bottom-up approach. Figure 5 below illustrates an example of an event that was automatically determined.

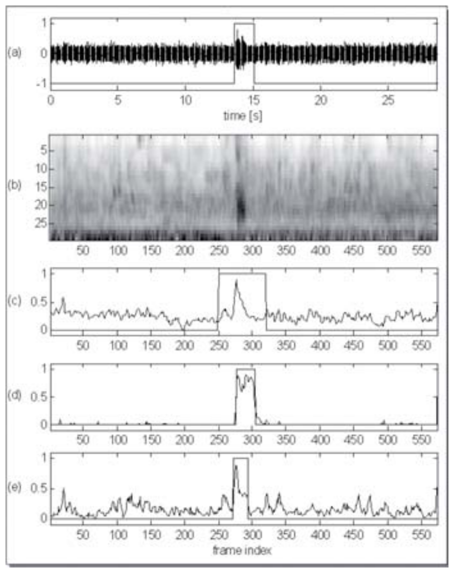


Fig. 5. Event detected by Normalized Auditory Attention Levels

- Unsupervised training for segmentation and modelling of background noise and known events performed by Thales. Here the idea is to continuously build a model of the environment, including background noise and known audio events, but also incorporating new audio events since an analysis of data from Rome has shown that 80% can be attributed to background (noise, speaking and 20% to normal events (trains, doors, music). Technically such processing is done in a multi-step procedure that includes vector quantization, decision trees for segmentation and Gaussian Mixture models for statistical modelling of the background. An example of a likelihood curve showing differences in background and audio event can be seen in Figure 6.
- Recognition of known events in the audio investigated by Brno University of Technology. The idea is to use detected events directly (for example to detect the brake sound) and make use of the uncertainty of such a system to estimate the segments with unusual sounds. The processing makes use of artificial neural networks and Viterbi decoding for event recognition. Its output can be seen in Figure 7. This approach was successfully demonstrated in on-line mode, processing the data from the audiovisual streaming server ●

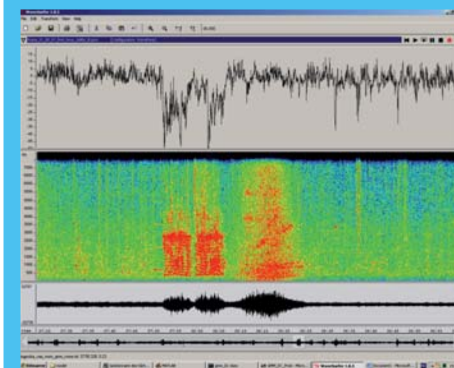


Fig. 6. Likelihood of a Gaussian Mixture Model showing difference between background and audio event

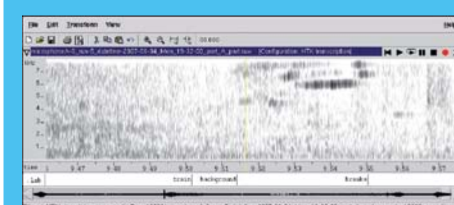


Fig. 7. Examples of known

CARETAKER consortium