# Multimedia Surveillance Systems

Rita Cucchiara
Dipartimento di Ingegneria dell' Informazione
University of Modena and Reggio Emilia
Italy
cucchiara.rita@unimore.it

## ABSTRACT

The integration of video technology and sensor networks constitutes the fundamental infrastructure for new generations of *multimedia surveillance systems*, where many different media streams, (audio, video, images, textual annotation, sensor signals) will concur to provide an automatic analysis of the controlled environment, and a real-time interpretation of the scene. New solutions can be devised to enlarge the view of traditional surveillance systems by means of distributed architectures of fixed and active cameras, to enhance the view with other sensed data, to explore multi-resolution views with zooming and omnidirectional cameras. Multimedia surveillance systems can be directly interfaced with biometric systems. Enlarge, enhanced and multiresolution views can handle face detection and recognition, face and speech correlation to support people identification VSSN05 is the third edition of the workshop, co-located at ACM Multimedia Conference, embracing research reports on video surveillance and sensor networks. This paper gives a short overview of the hot topics in multimedia surveillance systems and introduces some research activities currently engaged in the world and presented at VSSN05.

## 1. VIDEO SURVEILLANCE AND SENSOR NETWORKS

Nowadays, the majority of public opinion still conceives video surveillance systems as synonymous of CCTV systems: people imagines tens of cameras connected to tens of remote monitors, controlled by tens of bored and unheeding security employers; they should pay attention to people, vehicles, objects and suspicious situations to prevent crimes or disasters. In alternative, many believe that surveillance systems are storage platforms only, to memorize multimedia data (video, photos, wiretapped speech, textual information) available for human forensic experts to support investigations. This is true, and the value of these systems is undoubted. It is known, for instance, that the Italian Brigate Rosse terrorists, that killed in Bologna the welfare consultant of Italian Ministry Prof. Marco Biagi in 2002, where identified also thanks to the analysis of more than 52,000 hours of video recorded by surveillance cameras installed in the rail stations of Modena and Bologna. These systems are a concrete and profitable help to forensic investigations, although their potential capabilities are de-facto reduced by the limits of storage resource, and the consequent data compression and frame skipping. Instead, their effectiveness in prevention and real-time reactivity is still insufficient, since it is only in the hands of few human operators employed in the control that cannot manage all the huge amount of surveillance streams in real time.

Despite of the relevance of the current off-the-shelf surveillance systems, and their circumscribed role of mere support to human monitory, there is a world spread controversy about their use at large, connected with risks of privacy violations. The dichotomy safety vs. privacy was and is very debated in USA after September 11; there is an interesting paper of K.Bowyer discussing pros and cons and analyzing the risks of false claims in privacy violations [6]. In Europe, this discussion has been amplified after the terrorist attacks in Madrid in March 2004 and in the London underground in July 2005, that has been ineludible even if London is the city in the world with the highest number of cameras installed. Also in this case, the recorded videos have been constituted a valuable help in terrorists identification after the crime, but the systems were not capable to give an immediate alarm.

The society needs of the results of our research activities addressing new solutions in Video Surveillance and Sensor Networks; the tutelage of safety and security calls for new generations of multimedia surveillance systems, where computers will act not only as supporting platforms, but as the concrete core of real-time data comprehension process.

The scientific community is involved in theoretical studies on data understanding for surveillance since at least twenty years. All multimedia sources have been deeply explored: speech, image mosaics, sensor data, biometric data, handwritten texts and expecially video as stream to gather surveillance information. Many open problems are not solved yet but significant results have been carried out in many specific fields of video surveillance indeed, and consolidates theories have been transferred in commercial products too. A typical example is vehicle and traffic surveillance: systems for queue monitoring, accident and incident detection, tunnel monitoring and car plate recognition have been developed. For instance, the recent work of Kumar *et al.*, proposes a complete solution for vehicles detection and tracking by single camera, classification and recognition of vehicle behavior at intersections [24]. Similar approaches are proposed for human motion capture; in the field of single of few people tracking from single camera, the algorithms are now mature enough, if the acquiring conditions are not too prohibitive. Recent works propose algorithm for tracking multiple people and classify their posture both with 2D models [11] and 3D models [33]. There are still many open problems in tracking multiple people in non ideal conditions, in cluttered and unknown

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
*VSSN'05,* November 11, 2005, Singapore.
Copyright 2005 ACM 1-59593-242-9/05/0011 ...$5.00.

environment, with variable and unfavorable luminance conditions, for surveillance in indoor and outdoor spaces.

Although the problem of real-time visual content extraction is still partially unsolved for video streams from still cameras, most of the new research activities in surveillance are exploring larger dimensions: distributed video surveillance systems, heterogeneous video surveillance systems (with fixed, pan tilt zoom, omni-directional cameras), multi-spectral camera systems, surveillance system with multi-media streams (video, audio, sensor signals..), surveillance and biometric systems.

The topics, discussed in this $3^{rd}$ edition of ACM Workshop of Video Surveillance and Sensor Networks, reflect substantially the trend to explore new generations of multimedia surveillance systems. In this paper, a brief overview of the current research activities in multimedia surveillance systems is presented. The the future possibilities of integration of the surveiilance system and biometrics is discussed addressing in particula the problem of face detection. Some examples of experiments carried out at the Imagelab at hte University of Modena are presented. Finally a quick overview of the papers presented at the VSSN05 ACM workshop is given.

## 2. MULTIMEDIA SURVEILLANCE SYSTEMS

The adjective "multimedia" is normally referred to systems and services conceived for human end-users for accessing and using multimedia data, multimedia streams, multimedia content, multimedia interfaces in many different applications. Following this abstraction, a multimedia surveillance systems should be simply a surveillance systems capable of accomplishing the output of the task in a multimedia format; they should provide distilled video, images and sounds of the monitored environment, possibly annotated in an efficient and standard way, to improve further querying to surveillance stored data, or possibly transcoded in another media, such as text or animation. Instead, the concept of multimedia surveillance systems must be enlarged: it is not only a system capable to furnish multimedia data, but also to collect, process in real-time, correlate and handle multimedia data coming from different sources.

Multimedia surveillance systems are platform capable of including multi-modal video sources: distributed fixed cameras, pan tilt zoom, active and omni-directional cameras, multi-spectral cameras ( e.g. thermal cameras ). Multimedia surveillance systems can improve visual data with audio streams and information coming from other sensors. In large distributed environments, the exploitation of networks of small cooperative sensors (e.g. passive infrared wireless sensor (PIR) ) should improve dramatically the surveillance capability of few higher level sensors like the cameras. The research activity in distributed and multimedia surveillance systems is relatively recent; a pioneer project that defined a cooperative multi-sensor architecture was VSAM, for surveillance at the CMU campus [9].

The recent survey of Hu *et al.*, [19], reviews current developments of systems for motion detection and human identification, with information fusion from multiple cameras; as well, the review of Valera and Velastin gives a panorama of the so-called "intelligent distributes surveillance systems" typically oriented to handle multiple cameras in large environments [29]. In 2003, Brooks *et al.* defined a sensor network architecture for tracking and classification [7]. Many interesting research developments were presented in the two past editions of ACM Workshop of Video Surveillance and Sensor Networks [1], [2]. The NeST, [14], architecture was proposed to enlarge the view to the campus of USD with many distributed cameras: the software allow to show scene data archive and transcode them to permit a suitable and secure access with several etherogeneous mobile clients.

The WISE architecture [25] composed with wireless nodes and Panning cameras is presented together with algorithm to provide active surveillance of moving objects predicting their motion.

Multimedia surveillance systems "strech" the horizon of traditional video surveillance systems with different purposes:

- *Enlarge the view*: the Field of View (FoV) of a single fixed camera, or the Field of Regard of a single moving PTZ camera are limited in large environment; the straightforward goals of distributed systems with not overlapped cameras and possible networks are to allow an extended view of the scene;
- *Enhance the view*: even if the scene could be monitored by a single sensor, the adoption of redundant overlapped sensors can improve the understanding of the environment; overlapped cameras can give different views of the same target, cameras with audio or PIR sensors can disambiguate cluttered situations;
- *Explore multi-resolution views*: different media streams can be acquired from the the same point of view, with different granularity in order to have multi-resolution description of the scene and multiple level of abstraction. For instance, static camera views can be enriched by a Zoom camera, that gives a new view of a region of interest, generally exploited to recognize people capturing the faces or parts of the body.

The architecture of new generations of multimedia surveillance systems will be classified on the basis of different aspects:

1. the *stream granularity*: coarse-grain, as video clips, medium-grain as single images or single recorded audio stream, fine-grain such as the on-off signals of PIR sensors; have a different type of processed data, different semantic information content and affect real time performance;
2. the *data abstraction*: the data can be handled and stored simply in a very raw format (as pixel matrixes, compressed videos or key-frames), or post-processed ( as a segmented moving shape, a segmented voice), or identified objects (like a person with known identity, after a biometric analysis);
3. the *stream connectivity:* the streams can be completely uncorrelated each other, or can be taken in a loosely-coupled cooperative manner, such as videos of cameras with not overlapped field of view, or can be tightly-coupled each other, like the video streams acquired by synchronized overlapped cameras.

These aspects affect not only the real-time processing, but also the overall design of the data management system that will handle the results of the surveillance tasks. With multiple connected cameras, multiple sensors the problem of manage the wide amount of multimedia data will not be negligible; therefore, the choice and the organization of the semantic level of knowledge that is extracted and stored for further retrieving operation is particularly critical. For instance, the researches of Ebrahimi *et al.*, [13], explore approaches for standard production of MPEG-7 descriptions of surveillance-related data; Lopes *et al.* discusses the utility of MPEG-7 description of audio and video data for surveillance [26].

Concerning the stream connectivity, an important focus of the next multimedia surveillance systems will be the process synchronization, the cooperative integration, the scheduling and the planning of the multiple stream productions. Many research activities

**Figure 1: Dynamic mosaic with homography of two partially overlapped cameras. In the homography, note the correspondence between the bench and the person and their distorted shape flattered to the ground plane.**

(a) $C^1$ at frame #2881 (b) $C^2$ at frame #2881 (c) $C^3$ at frame #2881

**Figure 2: Examples of ImageLab system working with three cameras.**

currently addresses these issues: in 2003 Zhou *et al.* presented a master-slave architecture to detect humans for biometric recognition [35]; in the ACM VSSN04 Workshop Costello *et al* enlarged the analysis on scheduling algorithms for PTZ cameras used in people surveillance [10]. Many works of the 2005 edition of the workshop regards these topics.

A large effort should be combination of the multimedia data, beginning from the lowest level of abstraction. The image processing theories on image registration and the 3D geometry play an important role to reconstruct the 3D space with static and dynamic mosaics, homography and virtual models. The recent survey of Flusser and Zitova [36] gives an exhaustive presentation of many methods of image registration; An interesting approach to compute dynamic homography has been proposed in the ACM Workshop VS03 by Kang, Cohen and Medioni using a Tensor Voting- based algorithm. An example of homographic mosaic, computed in real time, to enlarge the view of surveillance system is shown in Fig.1, depicting some results in a public park of Reggio Emilia, Italy [12] . [21]. The computed homography and the mosaic with many overlapped cameras enables the surveillance system to track moving objects and people keeping consistency when objects are moving from a FoV to another.

As a case study, we report some experiments carried out with the ImageLab system in the Campus of Modena University to enlarge, enhance and exchange the views of video-surveillance systems with partially overlapped cameras. Fig. 2 shows the same person, viewed by three different cameras ($C^1$ and $C^2$ fixed and $C^3$ active PTZ camera), tracked concurrently by three modules, but with the same label. Cameras are not calibrated, but a training phase with a single walking person allows for an automated calibration [8]. The adoption of a reference walking people to reconstruct 3D information of a scene has been used both for traditional cameras [8] [22] and for omnidirectional cameras [15].

After the off-line training phase, a graph representation of consistency constraints of the whole system is synthesized and for each pair of overlapped cameras, epipole location is recovered exploiting parallax propriety of perspective images. The precise reconstruction of the 3D space allows a reliable tracking of the persons that are detected by the different camera systems independently. Each time new objects are detected in the overlapping area of a camera, i.e. at the camera handoff, they are projected in the overlapped camera plane by means of the homographic transformation. A Bayesian framework is exploited to provide information fusion, exploiting the knowledge coming from all the other cameras and solving possible disambiguations. If it is possible, unlabelled tracks are associated with tracks detected by other cameras and previously labelled. In this manner, each camera module can works autonomously, tracking the detected objects but with label consistent between modules (Fig. 4).

The Bayesian approach we have developed is particularly useful to disambiguate in case of groups of people: in Fig. 4.a two separate objects are detected in the left camera; both of their label are associated to the single blob detected by the right camera, that is not able to separate the elements by itself. The process can also correct errors of detection: in Fig. 4.c a person shape is over-segmented and split into two separate blobs (right camera, on the right) but

they are associated to a single label since they correspond to a single object detected in the other camera (see the person between two columns in the left image 4.c ). The system works with multiple cameras and with multiple targets as is depicted in the snapshot of Fig. 4.e with many people. Finally, the trajectories of all detected people can be annotated for surveillance purpose as in Fig. 4.g.

# 3. SURVEILLANCE AND BIOMETRIC SYSTEMS

Surveillance systems could have a twofold aim: they are developed to be used either in real-time to prevent disasters or crimes or/and to extract knowledge for a-posteriori investigation. In the former case, high reaction is required; in the latter, an high precision is preferred. Often these aspects are conflicting each other. Moreover, surveillance systems can have different final users; they could be exploited by Police experts in security offices; in addition, the market calls also for "intelligent" surveillance systems to be installed in the home and private offices for personal safety. In the first case, all present and past information on the environment, the detected objects and the people, should be available and possibly integrated with biometric data for identification purposes. In the second case, instead, for privacy and ethical issues, the identification of people and things (as number plate) is absolutely forbidden. Therefore, the possible synerhies of multimedia surveillance systems and biometrics systems will be deeply explored in the near future.

It is straightforward that media streams of surveillance systems could be further processed to determine the identity of detected people. Biometric data, as speech and face, could be matched against watching lists to identify unknown persons or to verify the claimed identity for authentication purposes. In the case of audio stream, speech recognition must be provided after a critical phase of filtering in order to discriminate voice from other noise and sounds; similarly in video streams people face can be recognized only after the non trivial process face detection. Some approaches search for face templates directly in raw images; other instead apply face detection after than people are detected (and tracked in the video). Complete surveys on human motion capture and people detection are [3] [16] [28].

Two recent surveys, of Yang *et al.* [32] and Hjelm and Low [17], collect a large number of proposals about face detection. Most of them are based on a skin color detection (like the one of Jones and Rehg [20]) followed by a face candidate validation by means of geometrical and topological constraints. If the face is detected in optimal pose and with a sufficient resolution, face recognition techniques achieve good results, as is detailed in the famous survey of face recognition of Zhao *et al* [34].

Enriching multimedia surveillance systems with biometric analysis to identify the detected persons is a goal of many recent proposals. For example, in [23] a system is proposed with a PTZ camera following, detecting and tracking faces that are recognized using FaceId software.

However, can biometric systems for face identification be coupled with surveillance systems in an effective way? Which is the required data granularity in terms of color and resolution that must be provided? Or, in an opposite way, which is the video resolution that can be adopted in surveillance systems that prevents people identification, in order not to contrast with the privacy issues?

About this last consideration, we would like to cite the old, but currently in use, UK standard about "privacy and forensic use of video material in CCTV systems". Given the Rotakin©standard (a human template) defined in 1994 by the Police Scientific Development Branch, (now HOSDB), a CCTV system with 625 rows that acquires a human template R of 160cm, is enabled for "monitoring and control" only if the FoV covers at least the 5% of the R, for "detection" if not less than the 10% of R is acquired, for "recognition" if not less than the 50% and for "identification" if not less than 120% [4]. "Monitor & control" means that an observer can determine the number, direction and speed of movement of people whose presence is known to him; "Detection" means that following an alert an observer can, after a search, ascertain with a high degree of certainty whether or not a person is visible in the pictures displayed to him. "Recognition", instead here means that viewers can say with a high degree of certainty whether or not the individual shown is the same as someone they have seen before. Eventually, "Identification" assumes that picture quality and detail should be sufficient to enable the identity of a subject to be established beyond reasonable doubts [4].

If we suppose a person walking standing, we can deduce how many pixels should be reserved for the face in an image to allow detection and identification for the UK standard. If we assume that for standard people the head in average is 1/8 of the whole body [1] , a frame is suitable for recognition and identification if the head's high is at least 39 and 93.5 pixels, respectively. Obviously the understanding ability of human is far beyond that of computer vision-based systems, but are the state-of-the-art of algorithms for face detection sufficiently robust to be acceptable for the aforementioned standard? To this aim, we tested one of the best assessed algorithm of face detection, i.e. the Viola-Jones approach [30]. This method proposes the adoption of the Haar transform to create pattern of interest and AdaBoost classifier to identify pixel patterns that can be considered as "faces". Table 1 shows an example of results of simple experiments provided at University of Modena with video streams from two different cameras, taken at four different resolutions with frontal and semi-profile (+/-15) poses. Ten versions of the sixteen situations have been replicated. Results using the algorithm of the OpenCV library [18], show that when the face is larger than 54x54 pixels the face detection is always correct. The correctness is acceptable with more than 48x48 pixels. For smaller sizes, the detection capability decreases. No face detection is possible for faces smaller than 40x40 pixels. These experiments confirm the effectiveness of this approach for many application with foreground faces. However, considering that the head's high is greater than the face size, it seems that the automatic algorithms cannot still satisfy the HOSDB standard of human capability of face detection. In [31] similar analysis has been carried out for face recognition and an accuracy of 90% has been achieved with a minimum resolution of 90x90 pixels only. It is worth noting that the Viola-Jones algorithm works on still images only, without exploit the temporal coherence and possible tracking information available from a video.

Experiments with head tracking show, instead, that the UK issues can be reached. In Table 2 some tests with different videos are reported with face detection achieved by head tracking [12]. This developed method integrates and improves the ideas proposed in [5] and [27]. The first searches for head template with both colour and gradient feature but the search space is limited to a neighbourhood of a predicted position. The problem of this solution is that it needs

[1]This assumption cames from the studies of Leonardo da Vinci that adopted this ratio, (formerly described by the Latin Vitruvio, and formerly by the Grecian Lipsia) in the illustrations of the book *Divinae Proportione* of Luca Pacioli published in 1509; actually, 1/8 ratio is correct only for people higher than 185 cm and is 1/6 for children. Moreover, since camera are normally installed in a high position, the ratio can be affected by the distortion due to perspective

| Video | Total Frames | Face h>54px | | 48px<Face h<54px | | 40px<Face h<48px | | Face h<40px | |
|---|---|---|---|---|---|---|---|---|---|
| | | N. frames | % detect. | N. frames | % detect. | N. frames | % detect. | N. frames | % detect. |
| $V_1$ frontal | 1005 | 585 | **100** | 100 | **100** | 120 | **60.00** | 200 | **0** |
| $V_2$ frontal | 750 | 416 | **100** | 100 | **91.00** | 134 | **42,53** | 100 | **0** |
| $V_3$ semi-profile | 1032 | 432 | **100** | 99 | **67,68** | 101 | **28,71** | 400 | **0** |
| $V_4$ semi-profile | 727 | 360 | **100** | 115 | **68,70** | 114 | **20,17** | 138 | **0** |

**Table 1: Experimental results using OpenCV face detector on faces of different sizes.**

| Video | N. frames | % Recogn. | Frontal | Profile | Horiz. View | Top View | Mean Face Height |
|---|---|---|---|---|---|---|---|
| V5 | 328 | 100% | 104 | 107 | 0 | 117 | 39 |
| V6 | 440 | 99% | 112 | 162 | 166 | 0 | 31 |

**Table 2: Performance of the face detection and tracking module**

a frame rate too high to make reliable predictions. Instead, [27] adopts a solution based on the elliptical Hough transform; it works at each frame and does not require any tracking nor prediction. A face colour histogram must be available as a model. Thus we use a supervised learning phase to compute a histogram of skin and hair colours when the head is selected in the first frames. It is done by taking the highest part of a tracked people when it enters in the scene or in case of camera hand-off. Thus, for each tracked object two different Hough transforms are computed: one gradient-based and one colour-based. The points belonging to the edges of the track vote for the first transform. Similarly, a point of the object votes for the colour-based transform if its colour has a non-zero value on the histogram of the saved head model. In this case, it votes for all the points inside an ellipse having the same size of the head and the actual pixel as the centre, and the rate is proportional to the model histogram value corresponding to the colour of the pixel. After that, the two transforms are normalized and cross-correlated to obtain a single map. The point with the higher value is chosen as the head centre. Table 2 shows some results. By exploiting people and head tracking, faces are detected in a correct way with a smaller head size and also with non frontal views. These techniques of face detection can be exploited in multimedia surveillance systems to extract the data useful for a possible people identification. If the aim of the system is security and forensics support, than the annotated faces can be inserted in a database, compared with the watching list and possibly identified. If the scope is crime prevention and improvement of the citizen safety and video surveillance systems are implemented for private market in accordance with the privacy lows, these approach could be exploited to hide identification features. Therefore the system can handle and store data with a higher level of abstraction, namely video streams of people without their identity. In the matter of fact, if the FoV is so large to have low resolution for the face, identification is not possible neither by human neither by automated procedures; if the resolution is sufficiently high that face can be detected, then, the head can be obscured, as shown in Fig. 3.

Robust techniques for real-time integration of biometric systems into multimedia surveillance systems will be available in the next future; the real-time extraction and selection of the visual, audio and biometric features will allow their obscuration, if required by the privacy low, and their annotation and interpretation, if not forbidden for security reasons.

# 4. MULTIMEDIA SURVEILLANCE SYSTEMS AT VSSN05

The works presented in the third edition of the ACM workshop VSSN05 embrace all the aspects of multimedia surveillance systems discussed in these pages.

## 4.1 Image and Video processing for Multimedia surveillance systems

Many research activities focus on the basic tasks of video processing from both single or multiple cameras. Working on the lowest level of data abstraction, i.e. on raw pixel data, many activities regard registration and calibration of images to enlarge the "view" by means of multi-camera systems. The work of **Heikkila and Pietikainen** proposes a fully automatic registration and mosaic of images, robust to rotation, scale and noise, for surveillance of wide areas. It is suitable for single keyframes and the background, since the SIFT feature detection, the feature matching, and the steps of warping and blending are still too computationally severe to be exploited for all the video frames in real-time. Nevertheles the idea with many poins belonging to the paper of **Horster et al.** proposes an automatic calibration of sets of cameras, using the same SIFT algorithm, and display for large multi-camera systems. Cameras do not have to share the same field of view. The paper of **Murino et al.** focuse on the automatic creation of initial background image annotating outlier trajectories of unknown moving objects. Methods of background suppression and frame difference are exploited to further extract Regions of Interests (ROI) for surveillance: in the paper of **Wang et al.** the position and the shape of detected ROI modify the output the display. In multimedia surveillance systems the video output is enriched augmenting the video parts that have an highest saliency. Finally, **Khalid and Naftel** propose an approach for unsupervised classification of extracted people trajectories with Self-Organizing Maps to give a semantic representation of the detected events.

## 4.2 Video audio and sensor networks to enlarge and enhance the view

Some works have been selected for VSSN05 described integrated architecture of audio, video and sensor networks for different applications. The work of **Wang te al.** describes a project of multi-camera architecture for airport surveillance, problems of sensors calibration, coordination are addressed and the definition of "soft biometry" is adopted to provide identification of the same person catched by different cameras in different poses. **Ardizzone et al.** illustrate a project of a multimedia surveillance systems for an archeological site with cameras and sensor networks. **Smeaton and McHugh** propose an approach of audio and video correlation for surveillance of indoor environments. **Prati et al.** present novel ap-

**Figure 3: Obscured faces**

proaches to enhance the "view" of surveillance systems with overlapped cameras enriched by PIR sensors that improve the people detection and disambiguate the uncertain detection in limit conditions (e.g. occluded views). The efficient and effective fusion of the information coming from different media streams is key point. **Atrey, Kankanhalli and Jain** in their paper extend the concept to information fusion with the definition of *information assimilation*: this process includes not only the real-time information fusion but also the integration with the past experience, represented by the surveillance information stored in the system. They describe a process to assimilate data from coarse and medium grain sensors, namely video and audio, and a probabilistic framework to discriminate concurring and contradictory evidences.

## 4.3 Explore multi-resolution views with PTZ and coordinated camera networks

As introduced previously, the integration of zoom cameras, overlapped cameras with different field of regards allows the acquisition and the correlation of new views of the same target. Calibration and coordination of these tightly-coupled sensors is mandatory. The paper of **Wren et al** describes an approach for calibrating a PTZ camera in large environment exploiting simple motion sensors. The network of small sensors are employed to accomplish a *functional calibration*: functional calibration is not a metric calibration but is a purposive process aiming to foveate targets, when required. Similarly, **Bagdanov et al.** exploit a formulation of the kinetic traveling salesperson problem to schedule the activity of a PTZ camera. The problem of surveillance camera scheduling is addressed at large by **Qureshi and Teropoulos**: they propose a simulation virtual environment where different control strategies can be tested and compared for wide field of view cameras and PTZ cameras. **Lim, Mittal and Davis** define a model to construct "task visibility intervals": they are exploited to indicate when an object can be visible form a sensor, which is the angle position of the camera and the duration of visibility. The goal is to schedule when a camera can acquire effective information of a target, for instance in people tracking to support face recognition. Finally, **Singh and Atrey** propose a new model called "coopetitive" using Model Predictive Control with cooperation and competition between sensors in order to swap the master-slave role. Examples to select the best views of target, e.g. intruders in a surveyed area, are shown.

# 5. CONCLUSIONS

In this paper, an overview of the recent research activities in multimedia surveillance systems has been presented. New generations of multimedia surveillance systems will cope with the problems of video, audio and sensor processing in order to real-time collect knowledge of the surveyed environment; they will deal with efficient annotation, display and access to surveillance multimedia data; they will address efficient policies for scheduling and controlling the sources of media streams loosely or tightly coupled. The final goal of these systems wil be to enlarge and enhance the view of the visible scene and explore new views with active cameras and other sensors. The papers presented in VSSN05 represent the state-of-the-art of this emerging research field.

The author would like to thanks Simone Calderara and Roberto Vezzani for the examples provided, and the Comune of Reggio Emilia, and the LAICA (Laboratorio di Ambient Intelligence per una Citta' Amica) project, for the images of the public parks.

# 6. REFERENCES

[1] *Proc. of ACM Workshop of Video Surveillance*, Nov 2003.

[2] *Proc. of the second ACM Workshop of Video Surveillance & Sensor Networks*, oct 2004.

[3] J. K. Aggarwal and Q. Cai. Human motion analysis: a review. *Computer Vision and Image Understanding*, 73(3):428–440, 1999.

[4] J. Aldrige and C. Gilbert. Testing on cctv perimeter surveillance systems. *In PSDB Publication*, (14), 1995.

[5] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, pages 232–237, 1998.

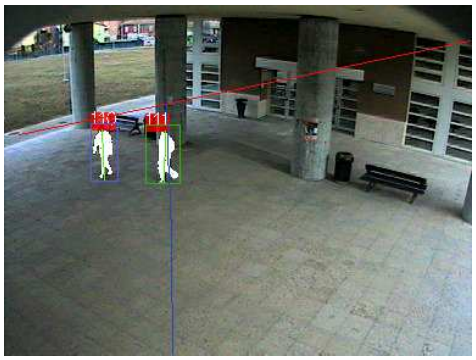
[6] K.W. Bowyer. Face recognition technology and the security versus privacy tradeoff. In *IEEE Technology and Society*, volume 1, pages 9–20, 2004.

[7] R.R. Brooks, P. Ramanathan, and A.M. Sayeed. Distributed target classification and tracking in sensor networks. In *Proc. of the IEEE*, volume 91, pages 1163 – 1171, 2003.

[8] S. Calderara, R. Vezzani, A Prati, and R. Cucchiara. Entry edge of field of view for multi camera tracking in distributed video surveillance. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2005.

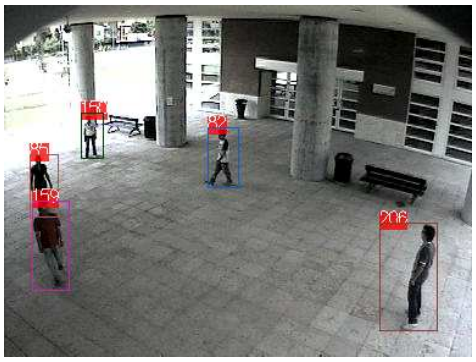
[9] R.T. Collins, A.J. Lipton, H. Fujiyoshi, and T. Kanade. Algorithms for cooperative multisensor surveillance. In *Proc. of the IEEE*, volume 89, pages 1456 – 1477, Oct. 2001.

[10] C.J Costello, C.P. Diehl, A. Banerjee, and H. Fisher. Scheduling an active camera to observe people. In *Proc of the ACM Workshop on Video Surveillance and Sensor Network*, pages 39–45, 2004.



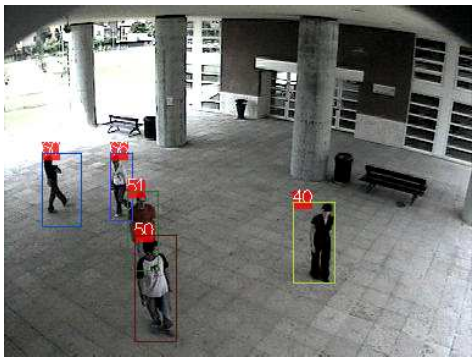
(a) Detection of disjoint two persons on $C^1$ at frame #432

(b) Disambiguation of grouped persons on $C^2$ at frame #432

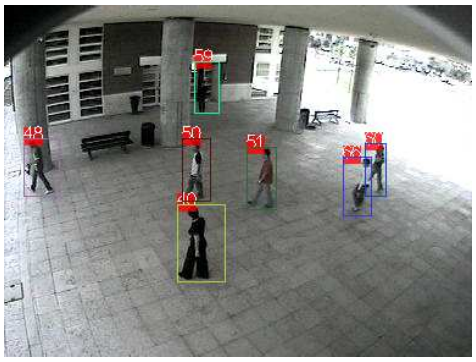


(c) Correctly segmented people in $C^1$

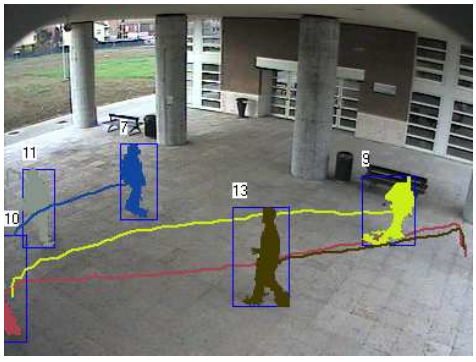
(d) Over segmented person in $C^2$



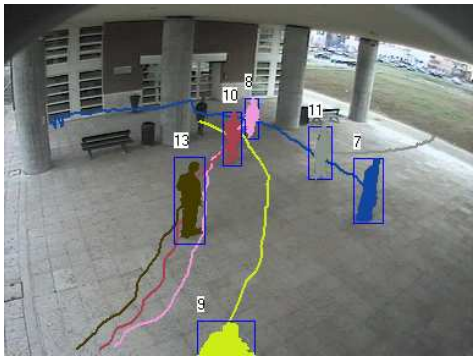
(e) Example of consistent labeling with many people ($C^1$)



(f) Example of consistent labeling with many people ($C^2$)



(g) Consistent trajectories extraction $C^1$



(h) Consistent trajectories extraction $C^2$

**Figure 4: Examples of system working on real cases with two cameras.**

[11] R. Cucchiara, C. Grana, A. Prati, and R. Vezzani. Probabilistic posture classification for indoor surveillance. *IEEE Trans. on Systems, Man, and Cybernetics - Part A*, 35(1):42–54, jan 2005.

[12] R. Cucchiara, A. Prati, and R. Vezzani. Ambient intelligence for security in public parks: the laica project. In *Proc. of IEE International Symposium on Imaging for Crime Detection and Prevention*, 2005.

[13] T. Ebrahimi, Y. Abdeljaoued, R. Figueras i Ventura, and O. Divorra Escoda. Mpeg-7 camera. In *Proc. of IEEE Int'l Conf. on Image Processing*, volume 3, pages 600–603, 2001.

[14] A. Fidaleo, H. Nguyen, and M. Trivedi. The network sensor tapestry(nest): A privacy enhanced software architecture for interactive analysis of data in video-sensor networks. In *Proc of the ACM Workshop on Video Surveillance and Sensor Network*, pages 46–53, 2004.

[15] T. Gandhi and M. Trivedi. Calibration of a reconfigurable array of omnidirectional cameras using a moving person. In *Proceedings of the ACM Workshop on Video surveillance & Sensor Networks*, pages 12–19, 2004.

[16] D. M. Gavrila. The visual analysis of human movement: a survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

[17] E. Hjelm and B.K. Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83(3):236–274, 2001.

[18] http://www.intel.com/research/mrl/research/opencv/. OpenCV library, Intel.

[19] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. on Systems, Man, and Cybernetics - Part C*, 34(3):334–352, August 2004.

[20] M.J. Jones and J.M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46:81–96, 2002.

[21] Jinman Kang, I. Cohen, and G. Medioni. Continuous tracking within and across camera streams. In *Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, volume 1, pages I–267 – I–272, 2003.

[22] S. Khan and M. Shah. Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(10):1355–1360, October 2003.

[23] Y.O. Kim, J. Paik, A. Jingu Heo Koschan, B. Abidi, and M. Abidi. Automatic face region tracking for highly accurate face recognition in unconstrained environments. *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS'03)*, pages 29–36, 2003.

[24] P. Kumar, S. Ranganath, Huang Weimin, and K. Sengupta. Framework for real-time behavior interpretation from traffic video. *IEEE Transactions on Intelligent Transportation Systems*, 6(1):43–53, March 2005.

[25] K-Y. Lam and C.K.H. Chiu. Adaptive visual object surveillance with continuously moving panning camera. In *Proc of the ACM Workshop on Video Surveillance and Sensor Network*, pages 29–38, 2004.

[26] R.J. Lopes, A.T. Lindsay, and D. Hutchison. The utility of mpeg-7 systems in audio-visual applications with multiple streams. *IEEE Trans. Circuits Systems for Video Technology*, 13(1):16–25, 2003.

[27] D. Maio and D. Maltoni. Real-time face location on gray-scale static images. *Pattern Recognition*, 33(9):1525–1539, sep 2000.

[28] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001.

[29] M. Valera and S.A. Velastin. Intelligent distributed surveillance systems: a review vision. In *Image and Signal Processing, IEE Proceedings*, volume 152, pages 192 – 204, April 2005.

[30] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2001.

[31] L. Wang, W. Hu, T. Tan, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. on Systems, Man and Cybernetics*, 3:334–352, 2004.

[32] M. Yang, D.J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

[33] T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(9), Sept. 2004.

[34] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 4(35):399–458, 2003.

[35] X Zhou, R.T Collins, T. Kanade, and P. Metes. A master-slave system to acquire biometric imagery of humans at distance. In *First ACM SIGMM Intl. workshop on Video surveillance*, pages 113–120, 2003.

[36] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, (21):977–1000, 2003.