Audio-Visual Event Recognition in Surveillance Video Sequences

Marco Cristani, *Member, IEEE*, Manuele Bicego, *Member, IEEE*, and Vittorio Murino, *Senior Member, IEEE*

***Abstract*—In the context of the automated surveillance field, automatic scene analysis and understanding systems typically consider only *visual* information, whereas other modalities, such as audio, are typically disregarded. This paper presents a new method able to integrate audio and visual information for scene analysis in a typical surveillance scenario, using only one camera and one monaural microphone. Visual information is analyzed by a standard visual background/foreground (BG/FG) modelling module, enhanced with a novelty detection stage and coupled with an audio BG/FG modelling scheme. These processes permit one to detect separate audio and visual patterns representing unusual unimodal events in a scene. The integration of audio and visual data is subsequently performed by exploiting the concept of synchrony between such events. The audio-visual (AV) association is carried out on-line and without need for training sequences, and is actually based on the computation of a characteristic feature called *audio-video concurrence* matrix, allowing one to detect and segment AV events, as well as to discriminate between them. Experimental tests involving classification and clustering of events show all the potentialities of the proposed approach, also in comparison with the results obtained by employing the single modalities and without considering the synchrony issue.**

***Index Terms*—Audio-visual analysis, automated surveillance, event classification and clustering, multimodal background modelling and foreground detection, multimodality, scene analysis.**

I. Introduction

The automatic monitoring of human activities has been of increasing importance in the last few years, thanks to its usefulness in the surveillance and protection of critical infrastructures and civil areas. This trend has amplified the interest of the scientific community in the field of video sequence analysis and, more generally, in the pattern recognition area [1]: the final aim is to design image-analysis systems that model and distinguish complex events and people activities like a human operator.

Typically, these systems often rely on a hierarchical framework: in the first phase, the raw data are processed in order to extract low-level information, which is subsequently processed by higher-level modules for scene understanding. In such a framework, an important low-level analysis is the so-called background modeling [2], [3], aimed at discriminating the expected information, namely, the background (BG), from the raw data to uniquely describe the current event, i.e., the foreground (FG).

In general, almost all human-activity recognition systems work mainly at the visual level only, but other information modalities can be easily available (e.g., audio) and used as complementary information to discover and explain interesting "activity patterns" in a scene. Computer Vision researchers have devoted their efforts to audio-visual (AV) data fusion in the video surveillance subfield only in the last few years (see Section II for a critical review of the related literature).

This paper aims to explore this research trend, proposing a novel strategy for activity analysis able to integrate audio and video information *at the feature level.* Video information is provided by a BG modeling system (based on a time-adaptive per-pixel mixture of Gaussian process [2]) able to model the background of a static scene while highlighting the foreground. This system is enhanced with a novelty-detection module aimed at detecting new objects appearing in a scene, thus allowing one to discriminate different FG entities. Monaural audio information is acquired by introducing the idea of FG audio events, i.e., unexpected audio patterns, which are detected automatically by modeling the audio background in an adaptive way. The adaptive video and audio modules work on-line and in parallel, so that, at each time step, they can detect separate audio and visual FG patterns in the scene.

On top of the unimodal processing stages, there is the core module, aimed at establishing a binding of audio and visual modalities, so that correlated audio and video cues can be aggregated and lead to the detection of AV events. This binding process is based on the notion of *synchrony* among the unimodal FG events occurring in the scene. This choice is motivated by the fact that the simultaneity is one of the most powerful cues available for determining whether two events define a single or multiple entities, as stated in early studies about AV synchrony resulting from cognitive science [4]. Moreover, psychophysical studies have shown that human attention focuses preferably on sensory information perceived coupled in time, suppressing those cues that are not [5]; particular importance has been devoted to the study of situations in which inputs arrive through two different sensory modalities (such as sight and sound) [6].

In our approach, the binding process is realized by building and on-line updating the so-called *audio-video concurrence* (AVC) matrix. This matrix permits one to detect significant nonoverlapped joint AV events and represents a clear and meaningful description of them. Such representation, built on-line and without the need for training sequences, is so effective as to allow one to accurately discriminate between different AV events by using simple classification or clustering techniques, like K-nearest neighbors (KNN) [7].

Manuscript received October 17, 2005; revised June 15, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Deepak S. Turaga.

M. Cristani and V. Murino are with the Dipartimento di Informatica, University of Verona, 37134 Verona, Italy (e-mail: cristanm@sci.univr.it; vittorio.murino@univr.it).

M. Bicego is with the DEIR, University of Sassari, 34-07100 Sassari, Italy (e-mail: bicego@uniss.it).

Digital Object Identifier 10.1109/TMM.2006.886263

In summary, the paper introduces several concepts related to multimodal scene analysis, faces the involved problems, and shows the potentialities and possible future directions of the research. The major contributions of this work are summarized in the following. We introduce: 1) an audio BG/FG modeling system coupled with a related video BG/FG module, working on line in an adaptive way and able to detect separate audio and visual foregrounds at each time instant; 2) a method for integrating audio and video information in order to discover multisensory FG patterns, thus potentially increasing the capabilities of surveillance systems; 3) a multimodal and multidimensional feature (i.e., the AVC matrix), based on the concurrence between audio and video patterns, which is proved to be expressive of the events occurring in an observed scene; 4) an AV fusion criterion embedded in a probabilistic framework working on-line, without the necessity for training sequences.

The rest of the paper is organized as follows. Section II reviews the AV fusion literature, clearly pointing out the main differences between the proposed approach and the state of the art. In Section III, the whole strategy is detailed, and experimental results are reported in Section IV. Finally, in Section V, conclusions are drawn and future perspectives are envisaged.

II. State of the Art of Audio-Visual Analysis

In the context of AV data fusion, it is possible to identify two principal research fields: the AV association, in which audio data are spatialized by using a microphone array (mainly devoted to tracking tasks), and the more general AV analysis, in which the audio signal is acquired by using only one microphone.

In the former, the typical scenario is a known environment (mostly indoor), augmented with fixed cameras and acoustic sensors. Here, a multimodal system locates moving sound sources (persons, robots) by utilizing the audio signal time delays among the microphones and the spatial trajectories performed by the objects [8], [9]. In [8], the tackled situation regards a conference room equipped with 32 omni-directional microphones and two stereo cameras; in the room, a multiobject 3-D tracking is performed. For the same environmental configuration, in [10] an audio source separation application is proposed. In another application [11], the audio information (consisting of footstep sounds) is used to detect a walking person among other moving objects by using a framework based on dynamic Bayes networks.

Other approaches based on the learning and inference of a graphical model can be found in [12], in which person tracking in an indoor environment is performed using video and audio cues provided by a camera and two microphones, respectively. In [13], a two-layer HMM framework is used to model predetermined individual and group multimodal meeting actions.

The second class of approaches employs only one microphone. In this case, audio spatialization is no more explicitly recoverable, so the AV binding must rely on other techniques. A well-known technique is canonical correlation analysis (CCA) [14], a statistical way of measuring linear relationships between two multidimensional random variables. In the AV context, the random variables are represented by the audio and video signals, i.e., spectral bands for the audio space and image pixels for the video one. CCA extracts a linear combination of a subset of pixels and a subset of bands that are maximally correlated. The fundamental problem of CCA-based approaches is the need for a large amount of data, which consequently leads to off-line applications in which the visual regions that emit sounds are constrained to be well-localized in the scene. Therefore, this method well behaves in the case of strongly supervised applications. A CCA-based approach is represented by FaceSync [15], an algorithm that measures the degree of synchronization between the video image of a face and the associated audio signal. A solution to the demand for a huge amount of data is proposed in [16], in which a presumed sparsity of the AV events is exploited.

Another class of inter-modal relationship detection methods is based on the maximization of the mutual information (MMI) between two sets of multivariate random variables. AV systems based on mutual information maximization are proposed in [17] and [18]. In [19], an information theoretical approach to modeling audio and video signals by using Markov chains is proposed in which the audio and video joint densities are estimated by using a set of training sequences. The methods based on MMI inherit the potentialities and drawbacks of CCA approaches: in [20], the equivalence between CCA and MMI under certain hypotheses on the underlying distributions has been shown.

The explicit detection of synchrony between audio and video represents another way to detect cross-modality relations, even if not so deeply investigated by the computer vision community in terms of localization aims. For example, for what regards the context of video surveillance, in the approach proposed in [21], audio and visual patterns are used to train an incrementally structured Hidden Markov Model in order to detect unusual AV events. Here, the audio patterns are formed by Mel-frequency cepstral coefficients from the raw audio signal, and the video patterns are composed of motion and color features from the moving blocks of each frame. The joining of the audio and visual features is performed by simply connecting both patterns. To the best of our knowledge, this is the first approach that deals with multimodality in the automated surveillance context.

Another research field in which audio-video analysis is largely exploited is video retrieval by content, in which the objects to be analyzed are typically entertainment sequences (movies, commercials, news, etc.). The ultimate goal is to enable users to retrieve the desired video clip from among massive amounts of heterogeneous visual data in a semantically meaningful and efficient manner. In this field, high-level concepts, such as video objects and events (spatio–temporal relations among objects) are exploited. The heterogeneity of the sequences considered requires the use of general high-level approaches, further heavily relying on automatic video annotation techniques [22], [23].

The proposed approach is different from the state of the art presented above for what concerns both the complexity of the considered data (except from the cited hidden Markov model approach [21], in relation of that we propose a comparative test in this paper) and the basic idea underlying the analysis performed. In our setting, AV sequences come from a video surveillance context, in which the camera is still (apart from small movements) and the audio comes directly from the scene being monitored, without any kind of control. Then, regarding the nature of the proposed approach, we studied an intuitive and accurate

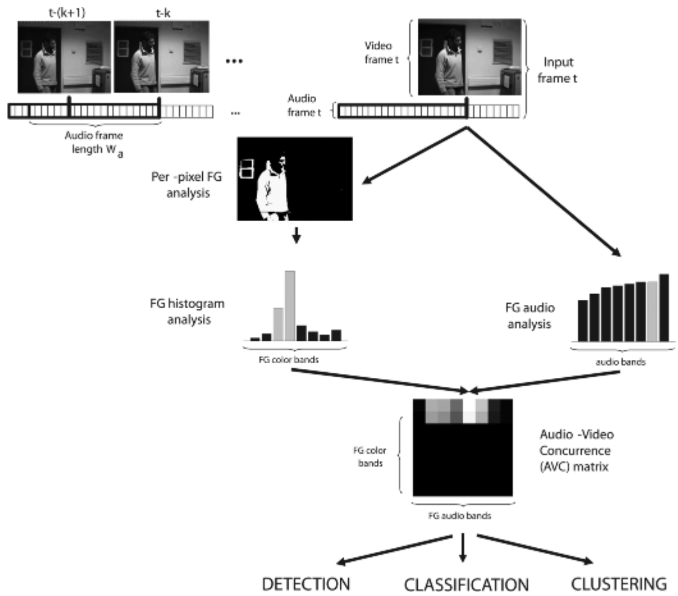

Fig. 1. Outline of the proposed system.

AV fusion criterion that does not require the formulation of any complex statistical model describing the relationships between audio and video information; it is working on line and without need for training sequences (as in [19], for example). In particular, the proposed method is heavily based on the concept of synchrony, which represents a well-motivated basic principle derived from psycho-physiological research.

Moreover, in our work we do not fit complex generative models with a large amount of concatenated audio-video data, as in [21] in a context of automated surveillance, or in [24] in the shot detection of football video sequences, where the shots are categorized into *play* and *break* [1] we prefer to process audio and video signals in order to discover multimodal association directly at the feature level, and then we use such features to perform clustering and classification tasks by applying simple algorithms.

III. The Proposed Method

A. Overview

The system is composed of several stages, starting with two separate audio and visual background modeling and foreground detection modules, as shown in Fig. 1.

For the visual channel, the model operates at two levels. The first is a typical time-adaptive per-pixel mixture of Gaussians model [2], able to identify the visual FG present in a scene. The second model works on the FG color histogram, and is able to detect different novel FG events. Despite the simple representation, this mixture model is able to characterize the appearance of FG data and to discriminate different FG objects.

Concerning the audio processing scheme, the concept of *audio* BG modeling[2] is introduced, capable to detect unexpected audio activities. A multiband frequency analysis is first carried out to characterize the monaural audio signal by extracting characteristic features from a parametric estimation of the power spectral density. The audio BG is then obtained by modeling such features related to each frequency band by using an adaptive mixture of Gaussians, so allowing one to detect, at each time step, a novel audio signal (e.g., a door that is closed, a ringing phone bell, etc., see Fig. 1). These modules work on-line, in parallel, and the outputs are the separate audio and video FG occurring in a scene at each time step.

AV association is subsequently developed by constructing the so-called AVC matrix, which encodes the degree of simultaneity of the audio and video FG patterns.

As assessed by psychophysical studies (see Section I), we assume that visual and audio FG that occur "simultaneously" are likely to be causally correlated. In particular, the FG contributions of each modality are collected at each time step, and then combined in the AVC matrix, whose i, j entry represents the importance of the audio FG energy localized in the ith audio subband *and* the FG *novel* appearance of a particular color range belonging to the jth FG histogram bin. This association is able to assess how much the inter-modal concurrence holds over time, permitting one to detect the most salient and permanent AV bindings. The resulting AVC matrix is therefore a multidimensional feature that, at each time step, summarizes and describes the AV activity occurring in the scene (see Fig. 1).

The high expressivity of such a feature allows one to effectively characterize and discriminate between such events, outperforming clustering and classification performances obtained by using individual modalities, as will be seen in the following.

The remainder of the section will give all the details of the proposed approach, starting from the basic time-adaptive mixture of Gaussians (Section III-B), and subsequently explaining how the video and audio channels are modeled (Sections III-C and III-D). Then, Section III-E provides details about the audio-video fusion, and Sections III-F and G contain the descriptions of how to perform AV event detection and discrimination, respectively.

B. The Time-Adaptive Mixture of Gaussians (TAPPMOG) Method

The TAPPMOG method is a probabilistic tool able to discover the deviance of a signal from the expected behavior in an on-line fashion, with the capability of adapting to a changing background.

In the general method [2], the temporal signal is modeled with a time-adaptive mixture of Gaussians with R components. The probability of observing the value $z^{(t)}$ at time t is given by

$$P(z^{(t)}) = \sum_{r=1}^{R} w_r^{(t)} \mathcal{N}\left(z^{(t)} | \mu_r^{(t)}, \sigma_r^{(t)}\right) \quad (1)$$

where $w_r^{(t)}$, $\mu_r^{(t)}$, and $\sigma_r^{(t)}$ a are the mixing coefficients, the mean, and the standard deviation, respectively, of the rth Gaussian of the mixture associated with the signal at time t. At each time instant t, the Gaussians are ranked in descending order using the w/σ value: the most ranked components represent the "expected" signal, or the background. Actually, the weight coefficient increases if the related Gaussian component

[1]Here the fusion of audio and visual features is performed *at the feature level* by simply concatenating them.

[2]A first-stage version appeared in [25].

models the observed signal for several consecutive frames; moreover, the σ value is smaller according to the stability of the signal value modeled. Stability and persistence make a signal "expected" or belonging to the background.

At each time instant, the ranked Gaussians are evaluated in descending order (with respect to w/σ) to find the first matching with the observation acquired (a match occurs if the value falls within 2.5σ of the mean of the component). If no match occurs, the least-ranked component (the least important) is discarded and replaced with a new Gaussian with the mean equal to the current value, a high variance, and a low mixing coefficient. If r_{hit} is the matched Gaussian component, the value $z^{(t)}$ is labeled as FG if

$$\sum_{r=1}^{r_{\text{hit}}} w_r^{(t)} > T \tag{2}$$

where T is a threshold representing the minimum portion of data that supports the "expected behavior." This concept can be easily understood by considering how the matching is determined. The Gaussian components are ordered in descending order by weight/variance: the best components (with a high weight and a low variance) are located in the first positions on the list. Summing the weights from the heaviest one up to the matched one gives us a quantity that is low (less than the threshold T) if the behavior is expected, i.e., modeled by the heaviest Gaussian components; otherwise, it is unexpected, i.e., modeled by lighter Gaussian components, ranked in the last positions in the mixture. We call the test in (2) the *FG test*, which is positive if the value is labeled as FG ($z^{(t)} \in FG$), and negative vice versa.

The equation that drives the evolution of the mixture's weight parameters is the following:

$$w_r^{(t)} = (1-\alpha)w_r^{(t-1)} + \alpha M^{(t)}, \quad 1 \leq r \leq R \tag{3}$$

where $M^{(t)}$ is 1 for the matched Gaussian (indexed by r_{hit}) and 0 for the others; the weights are renormalized at each iteration. Typically, the adaptive rate coefficient α remains fixed over time. The μ and σ of the matched Gaussian component are updated with the following formulas:

$$\mu_{r_{\text{hit}}}^{(t)} = (1-\rho)\mu_{r_{\text{hit}}}^{(t-1)} + \rho z^{(t)} \tag{4}$$

$$\sigma_{r_{\text{hit}}}^{2(t)} = (1-\rho)\sigma_{r_{\text{hit}}}^{2(t-1)} + \rho \left(z^{(t)} - \mu_{r_{\text{hit}}}^{(t)}\right)^T \left(z^{(t)} - \mu_{r_{\text{hit}}}^{(t)}\right) \tag{5}$$

where $\rho = \alpha \mathcal{N}\left(z^{(t)} | \mu_{r_{\text{hit}}}^{(t)}, \sigma_{r_{\text{hit}}}^{(t)}\right)$. The other parameters remain unchanged. It is worth noting that the higher the adaptive rate α, the faster the model is "adapted" to signal changes.

Concerning the initialization, the mixture of Gaussians is usually initialized with a bootstrap sequence in which a background situation is present. The length of the bootstrap sequence may change, starting from the shortest possible length of one value; actually, after the first value, the signal is modeled by one Gaussian component with the weight equal to one. Anyway, a longer bootstrap sequence ensures variance values of the Gaussian components that better model the noise present in the signal. In Section IV, we provide our initialization settings.

C. Visual Analysis

This section describes the visual module of the proposed system, which is able to detect atypical *visual* activity patterns. The designed method is composed of two parts: a standard per-pixel FG detection module and a histogram-based novelty detection module (see Fig. 1).

The former is a standard realization of the model described in (Section III-B), where each pixel signal $z_n^{(t)}$ is independently described by a TAPPMOG model: an unexpected valued pixel represents the visual per-pixel FG, $z_n^{(t)} \in FG$. Please note that all the mixtures' parameters are updated with a common fixed learning coefficient $\tilde{\alpha}$ and using a fixed value T as FG detection threshold: they are the same for audio and video channels.

The second module is a novelty detection system able to detect new objects appearing in a scene. This part is of basic importance for the proposed method, since the audio and visual pairing can be assessed if and only if a visual object appears in the scene (and remains FG) together with an audio FG signal: it is therefore fundamental to detect new objects appearing in the scene.

Toward this end, the idea is to compute at each time step the gray-level histogram of the sole FG pixels, which we called a *video foreground histogram* (VFGH). Each bin of the histogram, at time t, is denoted by $v_j^{(t)}$, where j varies from 1 to J, the number of bins. In practice, $v_j^{(t)}$ represents the quantity of pixels of the FG present in a scene at time t, with intensity values falling in the gray-level range j. Obviously, the accuracy of the description depends on the total number of bins J.

Then, we associate a TAPPMOG with each bin of the VFGH, looking for variations in the bins' values. When the number of foreground pixels significantly changes, obviously the related FG histogram also changes, and an occurring novel visual event can be inferred.

The probability of observing the value $v_j^{(t)}$, at time t, is modeled using a TAPPMOG

$$P\left(v_j^{(t)}\right) = \sum_{r=1}^{R} w_{(V,r,j)}^{(t)} \mathcal{N}\left(v_j^{(t)} | \mu_{(V,r,j)}^{(t)}, \sigma_{(V,r,j)}^{(t)}\right). \tag{6}$$

Defining u the matched Gaussian component, we can label the jth bin of the VFGH at the time step t as *visual FG value* if

$$\sum_{r=1}^{u} w_{(V,r,j)}^{(t)} > T. \tag{7}$$

This scheme permits one to detect both appearing and disappearing objects (an object is appearing in a scene when the bins' values suddenly increase, and it is disappearing when the bins' values decrease). Actually, we are interested only in appearing objects, since this represents the sole case in which AV synchrony is significant (a disappearing object, like a person that exits from the scene, should not be considered, as it does not belong to the scene anymore). Therefore, we disregard visual PG values deriving from negative variations in the foreground histogram bins, and consider only positive variations.

We are aware that the characterization based on the histogram causes some ambiguities (e.g., two equally colored objects are not distinguishable, even if the impact of this problem may be weakened by refining the number of bins), but this representation has the appealing characteristic of being invariant to the spatial localization of the FG. This characteristic is not recoverable by monitoring only the FG pixels directly.[3].

The computational complexity of this module is $O(N(R+1)+JR)$, where N is the number of pixels, R is the number of Gaussian components, the term 1 is due to the complexity of the PG histogram computation, and J is the number of PG bins.

D. Audio Analysis

The audio BG modeling module aims at extracting information from an audio signal acquired by a *single* microphone. In the literature, several taxonomies can be drawn in order to categorize the huge amount of available approaches. The "computational auditory scene analysis" (CASA) methods [27] translate psychoacoustics theories to automatically separate and classify sounds present in a specific environment by using signal processing techniques. The "computational auditory scene recognition" (CASR) approaches [28], [29] are aimed at the environment interpretation and do not analyze the different sound sources. More related to the statistical pattern recognition literature, a third class of approaches tried to merge "blind" statistical knowledge with biologically driven representations derived from the two previous fields, performing audio classification and segmentation tasks [30] and blind source separation [31], [32].

The approach presented in this paper falls into the third class. Roughly speaking, a multiband spectral analysis of the audio signal at the video frame rate is performed, extracting energy features from I frequency subbands, $a_1, a_2, \ldots, a_I$. More specifically, we subdivide the audio signal into overlapped temporal windows of fixed length W_a; each temporal window ends at the instant corresponding to the tth video frame[4] (see Fig. 1).

For each window, a parametric estimation of the power spectral density by the Yule–Walker autoregressive method [33] is performed; this method has been used by several time-series modeling approaches [34], [35], showing good performances for whatever audio window length. From this process, the energy samples [measured in decibels (dB)] $\left\{X^{(t)}(f_w)\right\}$, $w = 1, \ldots, W$ are derived, where f_w is the frequency expressed in Hertz, and the maximal frequency is $f_W = F_s/2$, where F_s is the sampling rate.

Subsequently, we introduce the *subband energy amount* (SEA), representing the histogram of the spectral energy, where each bin of the histogram, at time t, is denoted by $a_i^{(t)}$, $1 \leq i \leq I$. The SEA features have been chosen for their capability to discriminate between different sound events [28], [25], and because they can be easily computed at a high temporal rate, permitting one to discover unexpected audio behaviors for each channel at each time step.

Regarding the modeling of the time evolution of the SEA features, we assume that the energy over time at different frequency bands can provide independent information, as stated in [31]. Therefore, we instantiate one independent time-adaptive mixture of Gaussians (Section III-B) for each SEA channel. That is, the probability of observing the value $a_i^{(t)}$ at time t is modeled using a TAPPMOG

$$P\left(a_i^{(t)}\right) = \sum_{r=1}^{R} w_{(A,r,i)}^{(t)} \mathcal{N}\left(a_i^{(t)} | \mu_{(A,r,i)}, \sigma_{(A,r,i)}^{(t)}\right). \tag{8}$$

Let q be the Gaussian component matched when a new observation arrives; we can identify the SEA band value a_i as *audio FG value* if

$$\sum_{R=1}^{q} w_{(A,r,i)}^{(t)} > T \tag{9}$$

where the threshold T and the audio learning rate $\tilde{\alpha}$ are fixed and common parameters, equal to those used for the video channel.

The computational complexity of this module is $O(A+IR)$, where A is the (fixed) effort for the computation of the audio spectrum, I is the number of spectral channels, and R is the number of Gaussian components.

E. The AV Fusion

The audio and visual channels are now partitioned into different independent subspaces, the audio subbands $a_1, a_2, \ldots, a_I$, and the video FG histogram bins $v_1, v_2, \ldots, v_J$, respectively, in which independent unimodal FG values may occur. The basic idea is to find causal relations between each possible couple of audio and video bins at each time step t, on condition that both considered subspaces bring FG information.

Without loss of generality, let us consider the ith audio subspace and the jth video subspace; more specifically, let $a_i^{(t)}$ be the energy of the audio signal relative to the ith subband at the time step t, and let $v_i^{(t)}$ be the amount of FG pixels at the time step t in the scene that corresponds to the jth FG histogram bin.

Technically, we define a general *audio FG pattern* $A_i^{\left(t_{\text{init}}^A, t_{\text{end}}^A\right)}$ related to band i as the time interval when band a_i is foreground

$$A_i^{\left(t_{\text{init}}^A, t_{\text{end}}^A\right)} = \left[a_i^{\left(t_{\text{init}}^A\right)}, a_i^{\left(t_{\text{init}}^A+1\right)}, \ldots, a_i^{(t)}, \ldots, a_i^{\left(t_{\text{end}}^A\right)}\right] \tag{10}$$

where the interval $t_{\text{init}}^A, \ldots, t, \ldots, t_{ent}^A$ is such that $a_i^{(t)} \in FG$, $\forall\, t \in \left[t_{\text{init}}^A, t_{\text{end}}^A\right]$

In a very similar way, we can define the *video FG event* $V_j^{\left(t_{\text{init}}^V, t_{\text{end}}^V\right)}$, representing the interval time when the video foreground histogram band v_j is labeled as FG.

Given two FG patterns, we introduce the *potential relation interval (PRI)* as the time interval containing the possible overlapping of the audio and video patterns $A_i^{\left(t_{\text{init}}^A, t_{\text{end}}^A\right)}$ and $V_j^{\left(t_{\text{init}}^V, t_{\text{end}}^V\right)}$. If we define

$$t_{\text{init}}^{\text{AV}} = \max\left(t_{\text{init}}^A, t_{\text{init}}^V\right) \quad t_{\text{end}}^{\text{AV}} = \min\left(t_{\text{end}}^A, t_{\text{end}}^V\right)$$

[3]Actually, this is a simple way of detecting a novel FG without resorting to more sophisticated tracking approaches based on histograms [26], which will be subject of future work.

[4]In the following, we use a temporal indexing led by the *video* frame rate; therefore, the tth time step of the analysis is relative to the tth video frame.

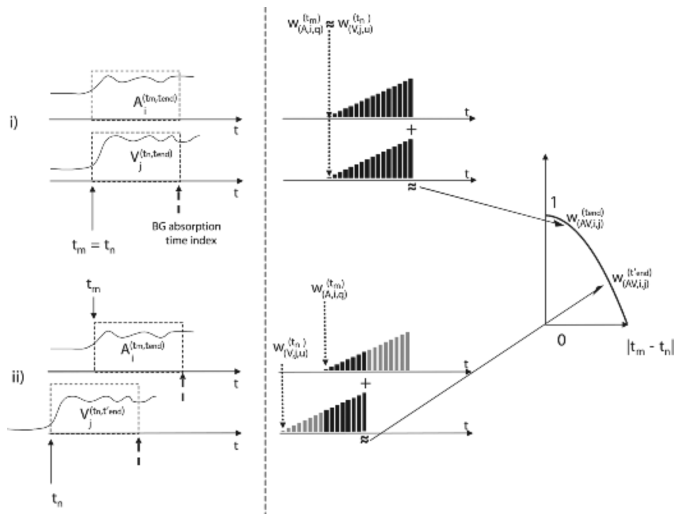

Fig. 2. Graphical definition of the AV coupling weight this value is able to distinguish different degrees of multimodal synchrony; in the left column, two cases of audio and video foreground patterns: (i) strongly synchronous and (ii) loosely synchronous. In the right column (on the left), the FG mixing coefficients of the Gaussian components that model the FG patterns. On the right, the behavior of the AV coupling weight: it is maximum when a complete overlapping of the FG patterns is present, and decreasing when the synchrony degree of the FG patterns diminishes.

the PRI can be described as

$$\mathrm{PRI}_{i,j}^{\left(A_{\mathrm{init}}^{\mathrm{AV}}, t_{\mathrm{end}}^{\mathrm{AV}}\right)} = \left[t_{\mathrm{init}}^{\mathrm{AV}}, t_{\mathrm{end}}^{\mathrm{AV}}\right] \tag{11}$$

where $t_{\mathrm{end}}^{\mathrm{AV}} > t_{\mathrm{init}}^{\mathrm{AV}}$. The $\mathrm{PRI}_{i,j}^{\left(t_{\mathrm{init}}^{\mathrm{AV}}, t_{\mathrm{end}}^{\mathrm{AV}}\right)}$ the time interval in which there is a concurrence represents between audio and video patterns, i.e., when the audio and video bands are synchronously FG.

Now we could define the AV coupling weight $w_{\mathrm{AV}}^{(t)}(i,j)$ as

$$w_{\mathrm{AV}}^{(t)}(i,j) = \begin{cases} \frac{w_{(A,i,q)}^{(t)} + w_{(V,j,u)}^{(t)}}{2}, & \text{if } t \in \mathrm{PRI}_{i,j}^{\left(t_{\mathrm{init}}^{\mathrm{AV}}, t_{\mathrm{end}}^{\mathrm{AV}}\right)} \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

where $w_{(A,i,q)}^{(t)}$ $\left(w_{(V,j,u)}^{(t)}\right)$ the weight of the Gaussian matched by the audio (video) band $a_i(t)$ $(v_j(t))$ in the audio (video) time-adaptive mixture of Gaussians model. If the information carried out by the audio channel i is in synchrony with the information carried out by the video channel j, then the patterns are correlated, and the AV coupling weight permits one to measure the strength of the AV association. A synthetic example is shown in Fig. 2.

We are now ready to introduce the main feature, namely, the AVC matrix. This matrix, of size $I \times J$, is able to accurately describe the AV history until time t: the i, j entry, at time t, is defined as

$$\mathrm{AVC}^{(t)}(i,j) = \sum_{t'=0}^{t} w_{\mathrm{AV}}^{(t')}(i,j). \tag{13}$$

At time $t = 0$, the matrix is empty. The AVC feature is computed on line, describes the audio-video synchrony from time 0 to t, and represents the core of the proposed approach. In Sections III-F and III-G, we shall see that AV event detection is derived directly from this feature, as well as a discriminative description of all AV events. Moreover, the AVC matrix, used in a surveillance context, permits the spatial localization of the audio foreground [36].

The computational complexity for the AVC feature calculation is obtained by simply adding the complexities of the audio and visual modules, and is linear in both the number of pixels and the number of Gaussian components used. This permits one to obtain high performances, as shown in Section IV.

F. AV Event Detection

The segmentation of the whole video sequence in AV events can be straightforwardly performed starting from the AVG matrix, Before describing how to segment the sequence, let us define an *audio video event (AVE)*: it occurs when an FG audio and an FG video are synchronously present in a scene. This can be detected by looking at the AVG matrix: if there is synchrony in the scene events, for some audio band a_i and some video band v_j, the AV coupling weight is nonzero. Therefore, an AVE is detected in the time interval $\left[t_{\mathrm{init}}^{\mathrm{AV}}, t_{\mathrm{end}}^{\mathrm{AV}}\right]$ if the following conditions hold simultaneously:[5]

1) $\mathrm{AVC}^{\left(t_{\mathrm{init}}^{\mathrm{AV}}-2\right)} - \mathrm{AVC}^{\left(t_{\mathrm{init}}^{\mathrm{AV}}-1\right)} = 0$ (no synchrony before $t_{\mathrm{init}}^{\mathrm{AV}}$).
2) $\forall\, t \in \left[t_{\mathrm{init}}^{\mathrm{AV}}, t_{\mathrm{end}}^{\mathrm{AV}}\right]$, $\mathrm{AVC}^{(t+1)} - \mathrm{AVC}^{(t)} \neq 0$ (synchrony during the event).
3) $\mathrm{AVC}^{\left(t_{\mathrm{end}}^{\mathrm{AV}}+1\right)} - \mathrm{AVC}^{\left(t_{\mathrm{end}}^{\mathrm{AV}}\right)} = 0$ (no synchrony after $t_{\mathrm{end}}^{\mathrm{AV}}$).

In other words, an audio-video event starts when the AVC matrix changes, and ends when the AVC matrix does not change anymore. Using this simple rule, we can segment on line the whole sequence in different K AV events AVE_k, $k = 1 \ldots K$, where each event is defined as

$$\mathrm{AVE}_j = \left[t_{\mathrm{init}}^{\mathrm{AV}}(k), t_{\mathrm{end}}^{\mathrm{AV}}(k)\right] \tag{14}$$

and $t_{\mathrm{init}}^{\mathrm{AV}}(k)$ $\left(t_{\mathrm{end}}^{\mathrm{AV}}(k)\right)$ indicates the initial (final) time step of the kth audio video event (see Fig. 3).

G. AV Event Discrimination

In the previous section, we have seen that the AVC matrix can be used to segment different AV events in the sequence. Nevertheless, this matrix could provide another useful information, since it contains also a rich description of the nature of an AV event; the description can be used for classifying the event. In detail, we propose to extract from the AVC matrix a feature, named *audio video description* (AVD), defined as

$$\mathrm{AVD}(\mathrm{AVE}_k) = \mathrm{AVC}^{\left(t_{\mathrm{end}}^{\mathrm{AV}}(k)\right)} - \mathrm{AVC}^{\left(t_{\mathrm{init}}^{\mathrm{AV}}(k)-1\right)}. \tag{15}$$

In simple words, this represents the AV information accumulated only during the event k. This matrix is then vectorized and directly used as a fingerprint vector for characterizing the AV event.

In the experimental part, we describe classification and clustering trials carried out on different audio video examples. In order to focus on the expressivity of such features, we performed clustering and classification using simple and

[5]Note that the following operations are computed among matrices: in particular, the relation of $\neq$ is valid if it holds for at least one matrix element i, j, while the relation $=$ is valid if it holds for all the elements.

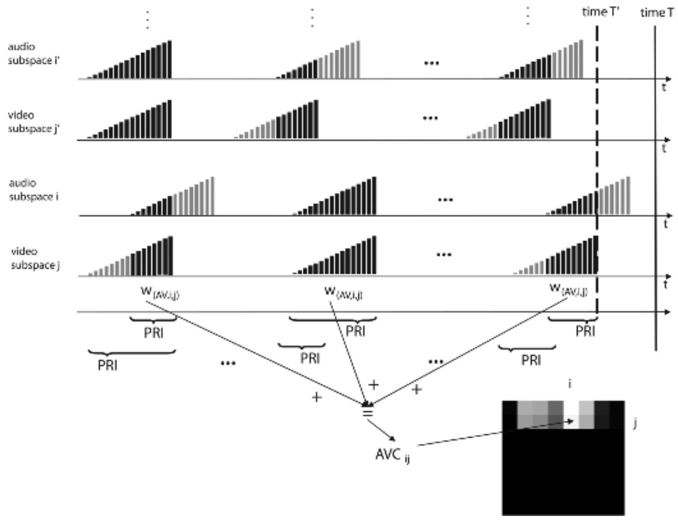

Fig. 3. Building process of the AVC matrix and AVE detection (the time indexing has been removed for clarity): considering the audio and video subspaces i and j (which stand for a_i and v_j, respectively), we sum in the position i, j of the AVC matrix all the AV coupling weights calculated until the time step T. The dashed line corresponds to the final step T' of an AVE. After this time step, no AV association holds between any subspaces i' and j', and the AVC matrix remains unchanged.

well-known methods, such as KNN for the classification, and hierarchical clustering.

IV. Experimental Results

In this section, we shall report various results obtained by applying our AV analysis to real video sequences. The aims of this section are 1) finding if the characterization of the AV events is meaningful (no over-segmentation or under-segmentation as compared with the segmentation performed by a human operator) and 2) testing if the features that describe the AV events are discriminant in terms of classification and clustering accuracy. In Section IV-A, we shall present the data set used and briefly discuss the roles of the parameters and their selection. In Section IV-B, we shall give an example of the computation of the AVC matrix, highlighting the key phases of the analysis. The remaining sections are devoted to showing the method performances in the: 1) detection (Section IV-C); 2) classification (Section IV-D); and 3) clustering (Section IV-E) processes.

A. Data Set and Parameter Setting

In order to test the proposed framework, we concentrate on different individual activities performed in an indoor environment, captured by using a standing camera and only one microphone. The activities (some shots are depicted in Fig. 4) are composed of basic actions, like entering the office, exiting the office, answering a phone call, talking, switching on/off the lights, and so on. Moreover, they are not overlapped, in the sense that the person appears in the scene, performs a set of basic actions and disappears, and then reappears later (with a time gap ranging from 0.5 to 10 s) to perform another sequence. The data gathering process was repeated in two sessions separated by three weeks. In each session, a further level of variability was due to the frequent change of clothes of the person in the video. The result was two long video sequences (more than 2 h overall). The sequences were captured by using a 320× 240 CCD camera, 20 frames per second. The audio signal was captured at 22050 Hz, and the samples were subdivided using temporal windows of length $W_a = 1$ s, and all the windows were overlapped by 70%.

Fig. 4. Pictures of the two sequences of activities.

For what concerns the number I of audio spectral subbands over which to calculate the SEA features and the number J of the FG histogram bins, we found that if we use $I = J = 8$, we have a good tradeoff between accuracy and low computational requirements, and obtain a near real-time computation of the AVC features (15 fps) using a PIII 500-MHz MATLAB implementation. Nevertheless, other experiments were performed using different I and J values, showing that this parameter is not crucial. We avoid using $I, J > 32$ because the curse-of-dimensionality problem may be incurred.

We considered the SEA of $I = 8$ equally subdivided subbands, in the range of $[0, 22050/4]$ Hz. A 3-component mixture of Gaussians was instantiated for each subband. This choice is not critical and can be suggested by opportune considerations about the complexity of the scene, especially in relation to the complexity of the background. Actually, three components are regarded as a reasonable choice, taking into account the possibility of a bimodal BG and one component for the foreground activity [2]. After some preliminary trials, the FG threshold T was set to 0.8, and the learning rate was set to $\alpha = 0.001$. The initial weights of the Gaussian components inserted in all the mixtures were fixed at $w_{\text{init}} = 0.001$.

For what concerns the video channel, we spatially subsample the video sequence by a factor of $\gamma = 4$ in order to speed up the computation of the per-pixel FG, and we use a 3-component mixture of Gaussians for each subsampled location. Then, we build the video FG histogram using $J = 8$ bins, obtained by equally partitioning the level of the FG gray interval $[0, 255]$ into eight intervals. Each of the corresponding FG histogram signals is modeled using again a 3-components mixture of Gaussians. Note that we use the same FG threshold T and the same learning rate for both the per-pixel FG detection and the video novelty detection. The only difference among the various mixture sets consists in the initial standard deviation σ_{init} with which the mixture components are initialized when a new Gaussian is added to the model (no match), due to the different ranges of variability of the values of the various subspaces. We noticed that these thresholds are important: too small an initial standard deviation means that we overfit the signal, introducing into the

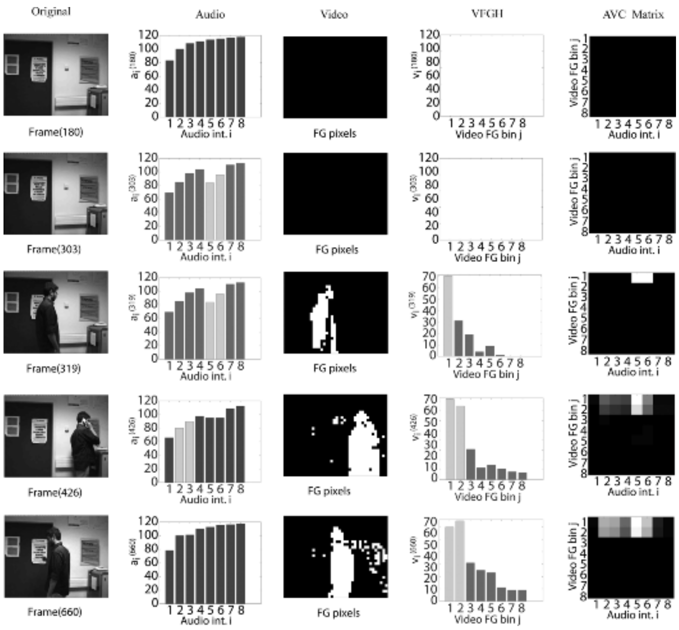

Fig. 5. AVC matrix computation steps for the "Phone" sequence.

mixture too many Gaussian components that are little different from one another, and consequently the resulting FG patterns are too numerous. On the other hand, too large an initial standard deviation means that we model with a unique Gaussian component a signal produced by different processes. In our experiments, we had that the range of variability of the SEA signal is about $[0, 150]$, for the pixel signal is $[0, 255]$ and for the VFGH signal is $[0, 320 * 240/\gamma^2 = 4800]$. After analyzing several different configurations (whose results are not shown here), we found that the best sensitivity parameters for the TAPPMOGs are $\sigma^A_{\text{init}} = 10$ for the audio subspace and $\sigma^P_{\text{init}} = 30$, and $\sigma^V_{\text{init}} = 50$ for the pixel and FG histogram subspaces, respectively.

B. An Illustrative Example

In order to understand the meaning of the AVC matrix, in this section we show its computation step by step, using a real sequence (namely, the "Phone" sequence) manually extracted from the experimental dataset. The sequence, 834 frames long, depicts a static scene in which a phone is located on the top of a piece of furniture; at a certain point, the phone rings, and a person comes and answers the call, concludes the conversation, and then goes out of the scene. In this sequence, the first 100 frames, depicting a background situation, are used as bootstrap sequence. This kind of initialization served also for the other following tasks.

A graphical representation of the AVC computation process is shown in Fig. 5. The salient points of the computation are

- frame 180: no FG patterns occur, neither audio nor video: the AVG matrix is empty;
- frame 303: the phone is ringing, as we see in the audio scheme depicting the SEA values, but no video FG is present; therefore, the AVC matrix remains empty;
- frame 319: the person comes and answer the call, then, the Video Foreground Histogram relative to the interval detects an FG video pattern that is concurrent with the audio FG pattern. Therefore, the starting point of an AV event is detected, and the AVG matrix shows some nonnull entries on the related AV coordinates;
- frame 426: the conversation continues, and consequently the AVG values increase;
- frame 660: the conversation ends, hence the FG patterns are over. The detection module communicates the end of the audio-video event, and the corresponding AVD feature can be computed.

The AVD represents the feature that will be used in the following classification and clustering tasks.

C. Detection Results

The sequence was segmented automatically into audio-video events using the definition presented in Section III-E. As ground truth, we asked a human operator to perform a segmentation of the two long sequences, highlighting human activities. Once the segmentation was performed, the 66 obtained segments were manually divided into six classes (situations), as follows.

1) *Make a call*: a person goes to the lab phone, dials a number, and makes a call.
2) *Receive a call*: the lab phone is ringing, a person goes to the phone and makes a conversation.
3) *First at work*: a person enters the lab, switches on the light, and walks in the room, without talking.
4) *Not first at work*: a person enters the lab with the light already switched on, walks in the room, and talks.
5) *Last at work*: a person exits from the lab, switching off the light without talking.
6) *Not last at work*: a person exits from the lab, leaving the light on and talking

Therefore, the two original long sequences were used as inputs to our system. The result of the automatic segmentation was optimal, in the sense that all the 66 events were identified as different. Moreover, our method was good as no over-segmentation or under-segmentation was performed; this was the primal element to be investigated in the paper. Further testing in which AV events occur overlapped is currently under study. Nevertheless, the goal of the work is to provide a feature-extraction technique able to ease the discrimination (clustering) and classification of different situations.

Anyway, note that this detection method applies actively the definition of AVD feature, and no further or more complex methods have to be used to separate the sequences. In other words, the detection step can be considered as a by-product of the proposed method. Anyway, if the events are overlapped and we need to separate them, a different detection approach has to be adopted.

D. Classification Results

We tested the classification accuracy for the 66 labeled audio-video events derived from the previous section in the four different scenarios listed below.

Scenario A—Situation 1 versus situation 2: making or receiving a phone call.

Scenario B—Situation 3 versus situation 4: entering an empty or nonempty lab.

TABLE I
LOO CLASSIFICATION ACCURACIES FOR THE FOUR DIFFERENT SCENARIOS

Scenario	Audio	Video	Audio-video
A	100.00%	86.35%	100.00%
B	60.87%	95.65%	95.65%
C	95.24%	85.71%	95.24%
D	62.12%	66.67%	89.39%

Scenario C—Situation 5 versus situation 6: exiting from an empty or nonempty lab.

Scenario D—Total problem: discrimination among all the six situations.

The classification was carried out using the nearest neighbor classifier with Euclidean distance [7], which represents the simplest classifier permitting one to understand the discriminative power of the proposed features. The classification accuracy was estimated using the leave one out (LOO) scheme [7]. In order to gain a deeper insight into the proposed method, we compared the proposed approach with the individual separate audio and video processings; in particular, the 66 audio and video FG patterns (see Section III-D and C), extracted in the same time intervals as the AVC features (i.e., per-channel summed over the PRI), were directly used as features to characterize the events. The classification accuracies for the three methods are presented in Table I. Just at a first look of the table, one can notice a general benefit from integrating audio and visual information: AV accuracies are the best results in all the experiments. Looking better at such figures, one can better figure out the outcome of the method, underlining some issues as follows.

- Scenario A is devoted to discriminating between making and receiving a phone call: clearly, most of the information is embedded in the audio part (when receiving a call, there is a ringing phone), whereas the visual part is really similar (going to the phone, hanging up, and talk). Actually, the audio signal itself is able to completely discriminate between these two events, whereas the video signal gets worse results. It is important to note that the AV integration does not inhibit the information brought in the audio part.
- Scenarios B and C are characterized by two similar AV situations. Regarding the audio part, there is a difference between talking in the lab and not talking, whereas regarding the video part there is a difference between switching on and off the lights. Actually, both single audio and video features yield good results, except for the audio score related to scenario B, which is rather low.
- Scenario D is the most complex and interesting. In this case, which involves six different classes, the integrated use of audio and video information permits one to drastically improve the classification accuracy by about 25%. The tasks are complicated, and only a proper integration of audio and visual information could lead to definitely satisfactory classification results.

In order to provide another comparative result, we considered a simple and straightforward multimodal approach, based on the concatenation of the audio and video FG patterns that we used previously as separate features, thus obtaining a multidimensional feature. This feature is similar in theory to those features used in [24] and [21]. We tested the classification accuracy of this method in the most difficult problem, i.e., the scenario D. In this case, we reached a classification accuracy equal to 82.28%: this is higher than the accuracies obtained by using the single modalities, but still lower than the accuracy reached by our approach: the synchrony approach adds information to the fusion consistently.

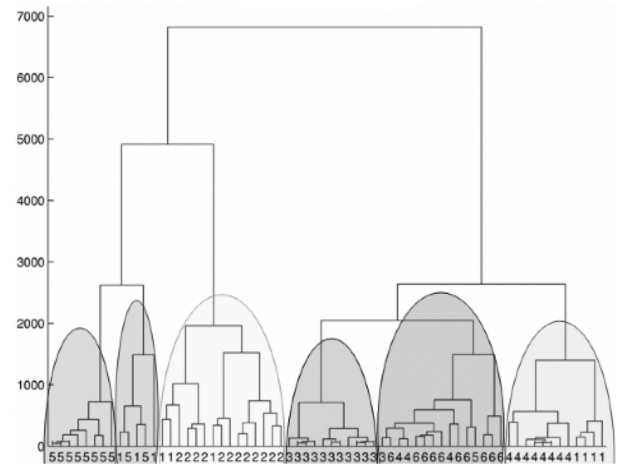

Fig. 6. Hierarchical clustering dendrogram.

E. Clustering Results

This last section reports results about clustering, in order to really detect patterns and natural groups of audio-video events. Given the automatically segmented dataset, we performed hierarchical clustering [37] using the Ward scheme, and we considered the Euclidean distance as the distance between elements. As in the classification task, we used a simple rule for performing clustering in order to define the expressivity of the AVD feature. We only set the number of clusters to six, and let the algorithm make the natural clusters. The resulting dendrogram is shown in Fig. 6, where the abscissas show the situation labels. Observing the dendrogram, we can see that the underlying structure of the dataset is satisfactorily represented, but, obviously, there are also some errors as the task is not easy. The most separated and best identified clusters are situations 2 and 3: they are characterized by clearly identifiable patterns (the ringing phone and the light on). Also, cluster 5 is quite well identified; little confusion is present between patterns in clusters 4 and 6. This is a difficult case: the discrimination between them lies only in the fact that in situation 4 the AV coupling occurs between the opening door and the person entering the scene (visual FG), and between the impulsive noise of the closing door and the speech signal (audio FG). Instead, in situation 6, the coupling between the closing door and the moving person is absent because the person went out of the scene producing no visual FG.

The clustering accuracy can be quantitatively assessed by computing the number of errors: a clustering error occurs if a pattern is assigned to a cluster in which the majority of the patterns belongs to another class. In this case, we obtain a clustering accuracy of 75.76%, which is a really satisfactory result, as compared with single modalities results: the clustering of the audio patterns obtained an accuracy of 53.03%, whereas the video pattern obtained 60.61%. The clustering operation using the simple concatenation of the audio and video patterns gave an accuracy of 64.44%. Also, this case points out the advantage of the fusion of audio-video information.

V. Conclusions

In this paper, a new method for characterizing audio visual events in a context of automated surveillance has been presented. Separate audio and video signals have been processed using two different adaptive modules aimed at considering audio and video information in a unique fashion, using only one camera and one microphone. Then, these two patterns have been integrated, exploiting the concept of synchrony in order to recognize audio-video events. The association has been realized by means of the AVC matrix, a feature that permits one to detect and segment AV events, and to discriminate among them. Experimental results on real sequences have shown promising results, in terms of both classification and clustering.

References

[1] R. C. T Kanade and A. Lipton, *IEEE Trans. Pattern Anal. Mach. Intell.—Special Issue on Video Surveillance*, vol. 22, no. 8, 2000.

[2] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR '99)*, 1999, vol. 2, pp. 246–252.

[3] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Proc. Int. Conf. Computer Vision*, 1999, pp. 255–261.

[4] T. Darrell, J. Fisher, and K. Wilson, Geometric and Statistical Approaches to Audiovisual Segmentation for Unthetered Interaction CLASS Project, 2002, Tech. Rep..

[5] E. Niebur, S. Hsiao, and K. Johnson, "Synchrony: A neuronal mechanism for attentional selection?," *Current Opinion Neurobiol.*, vol. 12, pp. 190–194, 2002.

[6] B. Stein and M. Meredith, *The Merging of the Senses*. Cambridge, MA: MIT Press, 1993.

[7] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[8] N. Checka and K. Wilson, Person Tracking Using Audio-Video Sensor Fusion MIT Artificial Intelligence Laboratory, Cambridge, MA, 2002, Tech. Rep..

[9] D. Zotkin, R. Duraiswami, and L. Davis, "Joint audio-visual tracking using particle filters," *EURASIP J. Appl. Signal Process.*, vol. 11, pp. 1154–1164, Nov. 2002.

[10] K. Wilson, N. Checka, D. Demirdjian, and T. Darrell, "Audio-video array source separation for perceptual user interfaces," in *Proc. Workshop on Perceptive User Interfaces*, 2001.

[11] X. Zou and B. Bhanu, "Tracking humans using multi-modal fusion," in *CVPR 05: Proc. 2005 Conf. Computer Vision and Pattern Recognition (CVPR '05)*, 2005.

[12] M. Beal, N. Jojic, and H. Attias, "A graphical model for audiovisual object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 828–836, Jul. 2003.

[13] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[14] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, Canonical Correlation Analysis an Overview With Application to Learning Methods Royal Holloway Univ. London, London, U.K., 2003, Tech. Rep. CSD-TR-03-02.

[15] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," *Proc. Neural Information Processing (NIPS 2000)*, 2000.

[16] X. Zou and B. Bhanu, "Pixels that sound," in *CVPR '05: Proc. 2005 Conf. Computer Vision and Pattern Recognition (CVPR '05)*, 2005, pp. 88–95.

[17] J. Fisher, III, T. Darrell, W. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2000, pp. 772–778.

[18] J. W. Fisher, III and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Trans. Multimedia*, vol. 6, no. 3, pp. 406–413, Jun. 2004.

[19] M. B. Cuadra, L. Cammoun, T. Butz, and J. C. O. Thiran, "From error probability to information theoretic (multi-modal) signal processing," *Signal Process.*, vol. 85, no. 5, pp. 875–902, 2005.

[20] J. Kay, "Feature discovery under contextual supervision using mutual information," in *Proc. Int. Joint Conf. Neural Networks*, 1992, vol. 4, pp. 79–84.

[21] D. Zhang, D. Gatica-Perez, and S. Bengio, "Semi-supervised adapted HMMs for unusual event detection," in *CVPR '05: Proc. 2005 Conf. Computer Vision and Pattern Recognition (CVPR '05)*, 2005, pp. 611–618.

[22] M. Petkovic and W. Jonker, *Content-Based Video Retrieval: A Database Perspective (Multimedia Systems and Applications)*. Berlin, Germany: Springer, 2003.

[23] S. Pfeiffer, R. Lienhart, and W. Efflsberg, "Scene determination based on video and audio features," *Multimedia Tools Appl.*, vol. 15, no. 1, pp. 59–81, 2001.

[24] M. Barnard, J. M. Odobez, and S. Bengio, "Multi-modal audio-visual event recognition for football analysis," in *Proc. IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, Sep. 2003, pp. 73–81.

[25] M. Cristani, M. Bicego, and V. Murino, "On-line adaptive background modelling for audio surveillance," in *Proc. Int. Conf. Pattern Recognition (ICPR 2004)*, 2004, pp. 399–402.

[26] M. Mason and Z. Duric, "Using histograms to detect and track objects in color video," in *Proc. 30th IEEE Applied Imagery Pattern Recognition Workshop (AIPR'01)*, Washington, DC, Oct. 2001, pp. 154–159.

[27] A. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. London, U.K.: MIT Press, 1990.

[28] V. Peltonen, "Computational Auditory Scene Recognition," Master's, Tampere Univ. Tech., Tampere, Finland, 2001.

[29] M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition," *Pattern Recognit. Lett.*, vol. 24, pp. 2895–2907, 2003.

[30] T. Zhang and C. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 4, pp. 441–457, May 2001.

[31] S. Roweis, "One microphone source separation," *Adv. Neur. Inform. Process. Syst.*, pp. 793–799, 2000.

[32] K. Hild, II, D. Erdogrnus, and J. Principe, "On-line minimum mutual information method for time-varying blind source separation," in *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA '01)*, 2001, pp. 126–131.

[33] S. Marple, *Digital Spectral Analysis*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1987.

[34] F. R. Bach and M. I. Jordan, "Learning graphical models for stationary time series," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2189–2199, Aug. 2004.

[35] P. Broersen, "Automatic spectral analysis with time series models," *IEEE Trans. Instrum. Meas.*, vol. 51, no. 2, pp. 211–216, Apr. 2002.

[36] M. Cristani, M. Bicego, and V. Murino, "Audio-video integration for background modelling," in *Proc. Eur. Conf. Computer Vision (ECCV 2004)*, 2004, pp. 202–213.

[37] A. Jain and R. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice-Hall, 1988.

Marco Cristani (M'06) received the Laurea degree in 2002 and the Ph.D. degree in 2006, both in computer science from the Università di Verona, Verona, Italy

He was a visiting Ph.D. student at the Computer Vision Lab, Institute for Robotics and Intelligent Systems School of Engineering (IRIS), University of Southern California, Los Angeles, in 2004–2005. He is now a Postdoctoral Fellow with the Dipartimento di Informatica, Università di Verona, working with the Vision, Image Processing and Sounds (VIPS) Lab. His main research interests include statistical pattern recognition, generative modeling via graphical models, and nonparametric data fusion techniques, with applications on surveillance, segmentation, and image and video retrieval. He is the author of several papers in the above subjects and a reviewer for several international conferences and journals.

Manuele Bicego (M'04) received the Laurea degree in 1999 and the Ph.D. degree in 2003, both in computer science from the Università di Verona, Verona, Italy

From 2000 to 2003, he was with the Vision, Image Processing and Sound (VIPS) Laboratory, Computer Science Department, Università di Verona, as a Supervisor of Research Activities in the areas of pattern recognition and image processing. In 2001 and 2006, he visited the Pattern Recognition and Telecommunications Laboratory, Instituto Superior Tecnico, Lisbon, Portugal. From 2004 to May 2005, he was a Postdoctorate Fellow at the University of Sassari, Sassari, Italy. He is currently a Researcher with the University of Sassari, and a member of the Computer Vision Laboratory. His research interests include statistical pattern recognition, electronic noses, hidden Markov models, video analysis, and biometrics. He is the author of several papers in the above subjects, published in international journals and conferences.

Dr. Bicego has been an Associate Editor of the *Electronic Letters on Computer Vision and Image Analysis* (ELCVIA) since 2004. He was a Guest Editor of the Special Issue on Similarity Based Pattern Recognition of *Pattern Recognition* in 2006. He served as a member of the scientific committee of three different international conferences in 2006, is a reviewer for several international conferences and journals, and a member of the IEEE Systems, Man, and Cybernetics society.

Vittorio Murino (SM'02) received the Laurea degree in electronic engineering in 1989 and the Ph.D. degree in electronic engineering and computer science in 1993, both from the University of Genoa, Genoa, Italy.

He is a Full Professor and Chairman of the Department of Computer Science, University of Verona. From 1993 to 1995, he was a Postdoctoral Fellow in the Signal Processing and Understanding Group, Department of Biophysical and Electronic Engineering, University of Genova, where he supervised of research activities on image processing for object recognition and pattern classification in underwater environments. From 1995 to 1998, he was an Assistant Professor of the Department of Mathematics and Computer Science, University of Udine, Udine, Italy. Since 1998, he has been with the University of Verona, where he founded and is responsible for the Vision, Image Processing, and Sound (VIPS) Laboratory. He is scientifically responsible for several national and European projects and is an Evaluator for the European Commission of research project proposals related to different scientific programmes and frameworks. His main research interests include computer vision and pattern recognition, probabilistic techniques for image and video processing, and methods for integrating graphics and vision. He is author or co-author of more than 150 papers published in refereed journals and international conferences.

Dr. Murino is a referee for several international journals, a member of the technical committees for several conferences (ECCV, ICPR, ICIP), and a member of the editorial board of *Pattern Recognition*, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, *Pattern Analysis and Applications*, and *Electronic Letters on Computer Vision and Image Analysis* (ELCVIA). He was the promotor and Guest Editor of four special issues of *Pattern Recognition* and is a Fellow of the IAPR.