Audio-Visual Foreground Extraction for Event Characterization

Marco Cristani, Manuele Bicego*, Vittorio Murino†

Abstract

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. The audio-visual association is performed on-line, by exploiting the concept of synchrony. Experimental tests carrying out classification and clustering of events show all the potentialities of the proposed approach, also in comparison with the results obtained by using the single modalities.

1. Introduction

The automatic monitoring of human activities has acquired an increased importance in the last years, due to its utility in the surveillance and protection of critical infrastructures and civil areas.

Typically, the surveillance 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 modelling [1], aimed at discriminating the expected information, namely, the background (BG), from those data to uniquely describe the current event, i.e., the foreground (FG).

In general, almost all human activity recognition systems work mainly at visual level only, but other information modalities can easily be available (e.g., audio), and used as complementary information to discover and explain interesting "activity patterns" in a scene. This paper wants to explore this research trend, proposing a novel strategy for activity analysis able to integrate audio and video information at the feature level.

The video information is provided by a BG modelling system, based on a time-adaptive per-pixel mixture of Gaussians process [1]. This system is enhanced with a novelty detection module aimed at detecting new objects appearing in the scene, thus allowing to discriminate different FG entities. The monaural audio information is acquired by introducing the idea of FG audio events, i.e. unexpected audio patterns, that are detected automatically by modelling in an adaptive way the audio background. This definition has been recently proposed by the authors in [2]. The adaptive video and audio modules work on-line and in parallel.

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 leading to the detection of *audio-visual* (AV) events. This binding process is based on the notion of *synchrony* between the unimodal FG events occurring in the scene. This choice is motivated from the fact that the simultaneity is one of the most powerful cues available for determining whether two events define a single or multiple objects, as stated in early studies about audio-visual synchrony coming from the cognitive science [3].

In our approach, the binding process is realized by building and on-line updating the so-called *Audio-Video Concurrence* (AVC) matrix. Such matrix permits to detect significant non-overlapped joint AV events and represents a clear and meaningful description of them. Such representation, built on line and without the need of training sequences, is so effective to allow to accurately discriminate between different AV events using simple classification or clustering techniques, like K-Nearest Neighbors (KNN) [4].

The rest of the paper is organized as follows. Section 2 reviews the AV fusion literature, clearly indicating the main differences between the proposed approach and the state of the art. In Section 3, the whole strategy is detailed, and experimental results are reported in Section 4. Finally, in Section 5, conclusions are drawn and future perspectives are envisaged.

*M. Bicego is with the DEIR, University of Sassari, via Torre Tonda, 34 - 07100 Sassari (Italy). Contacts: e-mail bicego@uniss.it, Tel: +39 079 2017321, Fax: + 39 079 2017312.

†M. Cristani and V. Murino are with the Dipartimento di Informatica, University of Verona, Strada le Grazie 15, 37134 Verona (Italy). Contacts: M. Cristani, e-mail cristanm@sci.univr.it, Tel: +39 045 8027072; V. Murino e-mail vittorio.murino@univr.it, Tel: +39 045 8027996, Fax: +39 045 8027068.

2. State of the art of the audio-visual analysis

In the context of audio-visual data fusion it is possible to individuate two principal research fields: the audio-visual association, in which audio data are spatialized using a microphone array (mainly devoted to tracking tasks), and the more general audio-visual analysis, in which the audio signal is acquired 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 by utilizing the audio signal time delays among the microphones and the spatial trajectories performed by the objects [5, 6]. In [5], the tackled situation regards a conference room equipped with 32 omnidirectional microphones and two stereo cameras, in which a multi-object 3D tracking is performed. In [6], the audio information (constituted by footstep sounds) is used to distinguish a walking person among other moving objects by using a framework based on dynamic Bayes nets.

The second class of approaches employs only one microphone. In this case, audio spatialization is no more explicitly recoverable, so the audio-visual binding must rely on other techniques. A well-known technique is the canonical correlation analysis (CCA), a statistical way of measuring linear relationships between two multidimensional random variables. A CCA-based approach is represented by FaceSync [7], an algorithm that measures the degree of synchronization between the video image of a face and the associated audio signal.

Another class of inter-modal relationship detection is based on the maximization of the mutual information (MMI) between two sets of multivariate random variables. The methods based on the MMI inherit the potentialities and the drawbacks of the CCA approaches: in [8], it has been shown the equivalence between CCA and MMI under certain hypotheses on the underlying distributions.

The explicit detection of synchrony between audio and video represents another way to detect cross-modality relations, even if not so deeply investigated from the computer vision community for what concerns localization aims. For example, in [9], audio and visual patterns are used to train an incrementally structured Hidden Markov Model in order to detect unusual AV events.

Another research field in which the audio-video analysis is largely exploited is the video retrieval by content, in which the objects to be analyzed are typically entertainment sequences (movies, commercials, news, etc.)[10].

The proposed approach is different with respect to those of the state of the art presented above from both what concerns the complexity of the considered data, and the basic idea underlying the analysis performed. In our setting, audio-video sequences come from a video surveillance context, in which the camera is still, apart small movements,

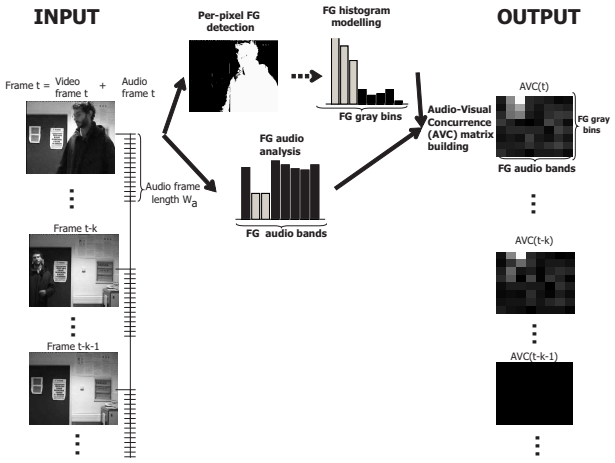

Figure 1. Outline of the proposed system.

and the audio comes directly from the scene being monitored, without any kind of control. Then, regarding to the nature of the proposed approach, we studied an intuitive and accurate audio-visual fusion criterion that do not require the formulation of any complex statistical model describing the relationships between audio and video information, working on line and without the need of training sequences. In particular, the proposed method is heavily based on the concept of synchrony, a well motivated basic principle derived from psycho-physiological research, also able to handle the localization issue.

Moreover, in our work we do not fit complex generative models with a large amount of concatenated audio-video data (like in [9]), preferring to process the audio and video signals in order to discover multimodal association directly at the feature level, and then, using such features to perform clustering and classification tasks by using simple algorithms.

3. The proposed method

3.1. Overview

The system is composed by several stages, starting with two separate audio and visual background modelling 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 [1], 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 modelling is introduced, capable to detect unexpected audio activities. A multi-band 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 modelling such features related to each frequency band using an adaptive mixture of Gaussians, see Fig.1). These modules work on-line, in parallel, and the output are the separate audio and video FG occurring in a scene at each time step.

Audio-visual association is subsequently developed by constructing the so-called *Audio-Video Concurrence (AVC)* matrix, which encodes the degree of simultaneity of the audio and the video FG patterns.
As assessed by psychophysical studies (see Sect. 1), we assume that visual and audio FG that occur "simultaneously" are likely causally correlated. The resulting AVC matrix represents a multidimensional feature that, at each time step, summarizes and describes the audio-visual activity being occurring in the scene (see Fig.1). The high expressivity of such feature allows to effectively characterize and discriminate between such events, outperforming clustering and classification performances obtained using the individual modalities.

3.2. The time-adaptive mixture of Gaussians method

In the general method [1], the temporal signal is modelled with a time-adaptive mixture of Gaussians with R components. The probability to observe the value $z^{(t)}$, at time t, is given by:

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

where $w_r^{(t)}$, $\mu_r^{(t)}$ and $\sigma_r^{(t)}$ are the mixing coefficients, the mean, and the standard deviation, respectively, of the r-th Gaussian of the mixture associated to the signal at time t. The Gaussians are ranked in descending order using the w/σ value: the most ranked components represent the "expected" signal, or the background.

At each time instant, the Gaussians are evaluated in descending order 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 last ranked component (the least important) is discarded and replaced with a new Gaussian with mean equal to the current value, high variance, and low mixing coefficient. If r_{hit} is the matched Gaussian component, the value $z^{(t)}$ is labelled as FG if $\sum_{r=1}^{r_{hit}} w_r^{(t)} > T$, where T is a threshold representing the minimum portion of the data that supports the "expected behavior". We call this test as *FG test*, that is positive if the value is labelled as FG ($z^{(t)} \in FG$), negative viceversa.

The equations that drive the evolution of the mixture parameters are the following :

$$w_r^{(t)} = (1-\alpha) w_r^{(t-1)} + \alpha M^{(t)}, 1 \leq r \leq R, \qquad (2)$$

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

$$\mu_{r_{hit}}^{(t)} = (1-\rho)\mu_{r_{hit}}^{(t-1)} + \rho z^{(t)} \qquad (3)$$

$$\sigma_{r_{hit}}^{2\,(t)} = (1-\rho)\sigma_{r_{hit}}^{2\,(t-1)} + \rho\left(z^{(t)} - \mu_{r_{hit}}^{(t)}\right)^T\left(z^{(t)} - \mu_{r_{hit}}^{(t)}\right) \qquad (4)$$

where $\rho = \alpha\mathcal{N}\left(z^{(t)} | \mu_{r_{hit}}^{(t)}, \sigma_{r_{hit}}^{(t)}\right)$. The other parameters remain unchanged.

3.3. Visual analysis

The designed method is composed by two parts. The former is a standard realization of the model explained in (Sec. 3.2), 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 mixtures' parameters are updated with a common fixed learning coefficient $\tilde{\alpha}$, and using a fixed value T as FG detection threshold, which are the same for audio and video channels.

The second module is a novelty detection system, able to detect when new objects appear in the scene. To this end, the idea is to compute at each time step the gray level histogram of the sole FG pixels, which we called *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 to each bin v_j of the VFGH, looking for variations of the bins' value. When the number of foreground pixels significantly change, also changes obviously the related FG histogram, and an occurring novel visual event can be inferred.
The probability to observe the value $v_j^{(t)}$, at time t, is modelled using a TAPPMOG:

$$P(v_j^{(t)}) = \sum_{r=1}^{R} w_{(V,r,j)}^{(t)} \mathcal{N}\left(v_j^{(t)} | \mu_{r,j}^{(t)}, \sigma_{r,j}^{(t)}\right) \qquad (5)$$

Defining u the matched Gaussian component, we can label the j-th bin of the VFGH at time step t as *visual FG value*, if $\sum_{r=1}^{u} w_{(V,r,j)}^{(t)} > T$.

This scheme permits to detect both appearing and disappearing objects (an object is appearing in the scene when bins suddenly increase their values, disappearing when bins values decrease). Actually, we are interested only in appearing objects, since this represents the sole case in which audio-visual 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). To this end, we disregard visual FG values deriving from negative variations of the foreground histogram bins, considering only the positive variations.

We are aware that the characterization based on the histogram leaves some ambiguities, but this representation has the appealing characteristic of being invariant to spatial localization of the FG (as in other audio-video analysis approaches). This characteristic is not recoverable by monitoring only the FG pixels directly. Actually, this is a simple way of detecting novel FG without resorting to more complex structured approaches, such as local histograms, edge histograms, shape histograms. These aspect will be envisaged in the future.

3.4. Audio analysis

The audio processing is composed by a multi-band spectral analysis of the audio signal at video frame rate. Here, we extract energy features from I frequency sub-bands, $a_1, a_2, \ldots, a_I$. More specifically, we subdivide the audio signal in overlapped temporal windows of fixed length W_a, in which each temporal window ends at the instant corresponding to the t-th video frame[1] (see Fig.1).
For each window, a parametric estimation of the power spectral density with the Yule-Walker Auto Regressive method is performed: this method has been used in several time series modelling approaches, showing good performances whatever audio window length is used. From this process, the energy samples (measured in decibel, dB) $\{X^{(t)}(f_w)\},\ w = 1, \ldots, W$ are obtained, where f_w is the frequency expressed in Hz, and the maximal frequency is $f_W = F_s/2$, with F_s 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 with $a_i^{(t)},\quad 1 \leq i \leq I$. The SEA features have been chosen for their capability to discriminate between different sound events , and because they can be easily computed at an elevate temporal rate, permitting to discover unexpected audio behaviors for each channel at each time step.

Regarding the modelling of the time evolution of the SEA features, is a proved assumption that the energy during time at different frequency bands can transport independent information.Therefore, we instantiate one independent time-adaptive mixture of Gaussians (Sec. 3.2) for each SEA channel. That is, the probability to observe the value $a_i^{(t)}$, at time t, is modelled using a TAPPMOG:

$$P(a_i^{(t)}) = \sum_{r=1}^{R} w_{(A,r,i)}^{(t)} \mathcal{N}\left(a_i^{(t)} | \mu_{r,i}^{(t)}, \sigma_{r,i}^{(t)}\right) \qquad (6)$$

Let q be the Gaussian component matched when new observation arrives, we can identify the SEA band value a_i as *audio FG value*, if $\sum_{r=1}^{q} w_{(A,r,i)}^{(t)} > T$, where the threshold T and the audio learning rate $\tilde{\alpha}$ are fixed and common parameters, equal to those used for the video channel.

3.5. The Audio-Visual fusion

The audio and visual channels are now partitioned in 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 leading idea is to find causal relations among each possible couple of audio and video bins at each time step t, with the condition that both considered subspaces bring FG information.
Without loss of generality, let's consider the $i-$th audio subspace and $j-$th video subspace; more specifically, let $a_i^{(t)}$ be the energy of the audio signal relative to the $i-$th subband at time step t and $v_j^{(t)}$ the amount of FG pixels at time step t in the scene (that correspond to the $j-$th FG histogram bin).

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

$$A_i^{(t_{init}^A, t_{end}^A)} = \left[a_i^{(t_{init}^A)}, a_i^{(t_{init}^A+1)}, \ldots, a_i^{(t)}, \ldots, a_i^{(t_{end}^A)}\right] \qquad (7)$$

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

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

Given two FG patterns, we introduce the *Potential Relation Interval* as the time interval containing the possible overlapping of the audio and video patterns $A_i^{(t_{init}^A, t_{end}^A)}$ and $V_j^{(t_{init}^V, t_{end}^V)}$. Defining

$$t_{init}^{AV} = \max(t_{init}^A, t_{init}^V) \qquad t_{end}^{AV} = \min(t_{end}^A, t_{end}^V)$$

then the Potential Relation Interval could be described as

$$PRI_{i,j}^{(t_{init}^{AV}, t_{end}^{AV})} = [t_{init}^{AV}, t_{end}^{AV}] \qquad (8)$$

where $t_{end}^{AV} > t_{init}^{AV}$. The $PRI_{i,j}^{(t_{init}^{AV}, t_{end}^{AV})}$ represents the time interval in which there is a concurrence between audio and

[1] In the following, we use a temporal indexing leaded by the *video* frame rate; therefore, the $t-$th time step of the analysis is relative to the $t-$th video frame.

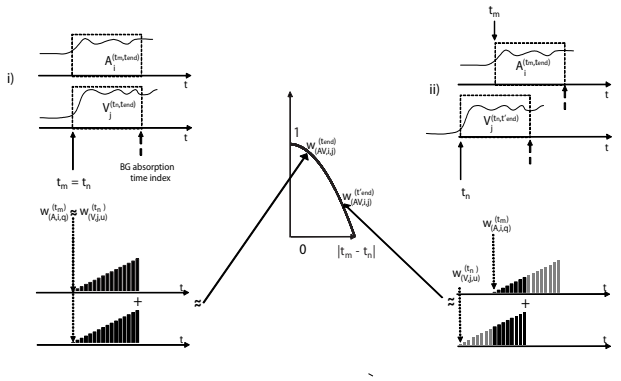

Figure 2. Graphical definition of the AV coupling weight: this value is able to distinguish different degrees of multimodal synchrony; on the left column, two cases of audio and video foreground patterns (i) strongly synchronous (ii) loosely synchronous. On the right column (on the left), the FG mixing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

video patterns, i.e., when the audio and video bands are synchronously FG.

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

$$w_{AV}^{(t)}(i,j) = \begin{cases} \dfrac{w_{(A,i,q)}^{(t)} + w_{(V,j,u)}^{(t)}}{2} & \text{if } t \in PRI_{i,j}^{(t_{init}^{AV}, t_{end}^{AV})} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where $w_{(A,i,q)}^{(t)}$ ($w_{(V,j,u)}^{(t)}$) is 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 synchronous with respect to the information carried out by the video channel j, then the patterns are correlated, and the AV coupling weight permits to measure the strength of the AV association. A synthetical example is shown in Fig. 2: on the left column, we see two cases of audio and video foreground patterns that (i) are strongly synchronous and (ii) are loosely synchronous. The area placed in the dashed-line box indicates a FG pattern, whose initial instant is indicated with a solid arrow, and the relative BG absorption is pointed out with a dashed arrow. On the right column (on the left), the mixing coefficients of the Gaussian components that model the FG patterns are indicated. One can notice that these coefficients (the first of them indicated by a dotted line) are proportional with the time spent by the related Gaussian components to model the FG values: the FG mixing coefficients increase (if the pattern is always modelled by the same component) until the FG test (explained in Sec. 3.2) is negative, then, such value is labelled as background. The AV coupling weights are built using the unimodal FG mixing coefficients in the related Potential Rela-

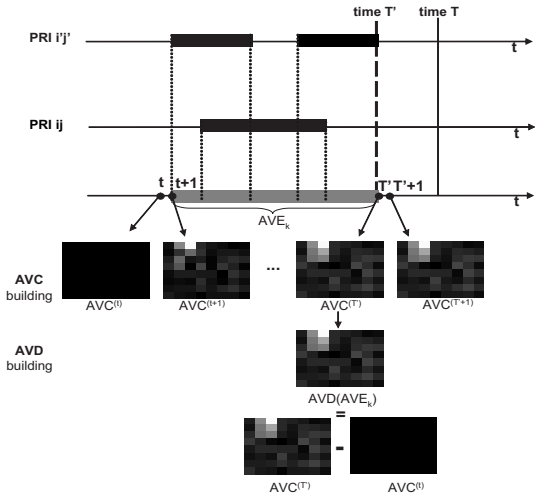

Figure 3. Building process of the AVC and AVD matrices: considering overlapping PRI's relative to audio and video subspaces i and j (on the bottom) and i' and j' (on the top), that form the k-th AVE, we can see how how the AVC is calculated until the time step T. The dashed line corresponds to the final step T' of the AVE. After that time step, no AV association holds among any subspaces, therefore AVC matrix remains unchanged and the AVD matrix relative to the AVE k is calculated.

tion Interval (indicated as black bars in the picture). As one can observe (on the right), this value is maximum when a complete overlapping of the FG patterns is present, while a strong decreasing appears when the synchrony degree of the FG patterns diminishes.

Now, we are ready to introduce the main feature, namely the *Audio-Video Concurrence (AVC)* matrix. This matrix, of size $I \times J$, is able to accurately describe the audio-visual history until time t: the i,j entry, at time t, is defined as:

$$AVC^{(t)}(i,j) = \sum_{t'=0}^{t} w_{AV}^{(t')}(i,j) \tag{10}$$

At time $t = 0$, this 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. We will see in the next sections that AV event detection is directly derived from this feature, as well as a discriminative description of the AV events.

3.6. Audio-visual event detection

The segmentation of the whole video sequence in audio-visual events can be straightforwardly performed starting from the AVC matrix. Before describing how to segment it, let us define an *Audio Video Event (AVE)*: it occurs when a FG audio and a FG video are synchronously present in the scene. This can be detected by looking at the AVC matrix: if there is synchrony in the scene events, for some audio band a_i and video band v_j, the AV coupling weight is non

zero. Therefore, an AVE is detected in the time interval $[t^{AV}_{init},\ t^{AV}_{end}]$ if the following conditions hold contemporaneously [2]:

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

In other words, an audio-video event starts when the AVC matrix changes, and terminates 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, $k = 1 \cdots K$, where each event is defined as

$$AVE_k = [t^{AV}_{init}(k),\ t^{AV}_{end}(k)] \tag{11}$$

and $t^{AV}_{init}(k)$ ($t^{AV}_{end}(k)$) indicates the initial (final) time step of the k-th audio video event (see Fig.3).

3.7. Audio-visual event discrimination

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

$$AVD(AVE_k) = AVC^{(t^{AV}_{end}(k))} - AVC^{(t^{AV}_{init}(k)-1)} \tag{12}$$

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.

4. Experimental Results

In this section, we will show various results obtained by applying our AV analysis to real video sequences. The aims of this section are multiple: 1) finding if the characterization of the AV events is meaningful (no over-segmentation or under-segmentation respect to a segmentation performed by a human operator); 2) testing if the features that describe the AV events are discriminant with respect to classification and clustering tasks. In Sec. 4.1, we will present the data set used, and we briefly discuss the role of the parameters and their selection. The remaining sections are devoted to show the method performances in the (i) detection (Sect. 4.2), (ii) classification (Sect. 4.3), and (iii) clustering (Sect. 4.4).

[2] 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.

4.1. Data set and parameter setting

We concentrate on different individual activities performed in an indoor environment, captured using a standing camera and only a single microphone. The activities (some shots are depicted in Fig. 4) are composed by 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, reappearing later (with time gap varying from 0.5 sec to 10 sec) to perform another sequence. The data gathering process was repeated in two separate sessions with three weeks of distance between them. 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 2 long video sequences (more than 2 hours overall). The sequences have

Figure 4. Some pictures of the two activities sequences.

been captured using a 320×240 CCD camera, 20 frames per second. The audio signal is captured at 22050 Hz, and the samples are subdivided using temporal windows with length $W_a = 1$ s, and all the windows are overlapped of 70%.

For what concerns the number I of audio spectral subbands over which calculating the SEA features and the number J of the FG histogram bins, we find that using $I = J = 8$ we have a good compromise between accuracy and low computational requirements. In specific, we have considered the SEA of $I = 8$ equally subdivided subbands, in the range of $[0, 22050/4]$Hz. A 3-components mixture of Gaussians have been instantiated for each subband. The FG threshold T has been set to 0.8, and the learning rate is set to $\alpha = 0.001$. For what concerns the video channel, we spatially sub-sample the video sequence by a factor of 4, in order to speed up the computation of the per-pixel FG, and we use a 3-components 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 FG gray interval [0,255] in 8 intervals. Each of the corresponding FG histogram signals is modelled using again 3-components mixture of Gaussians.

4.2. Detection results

The sequence has been segmented automatically in audio-video events using the definition presented in

Sec. 3.5. 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 classified in 6 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 into the lab, switch on the light, and walks in the room, without talking.
4) *Not first at work*: a person enters into 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 original sequences were used as input 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 in the sense that no over-segmentation or under-segmentation was obtained, primal element to be investigated in this paper. Further testing in which the AV events will occur overlapped is actually under exploration.

4.3. Classification results

We tested the classification accuracy with the 66 labelled audio video events derived from the previous part, in four different scenarios, listed below:
Scenario A - Situation 1 vs situation 2: making or receiving a phone call.
Scenario B - Situation 3 vs situation 4: entering in an empty or non empty lab.
Scenario C - Situation 5 vs situation 6: exiting from an empty or a non empty lab.
Scenario D - Total problem: discrimination between all the six situations.

The classification accuracy was computed using the Nearest Neighbor classifier with Euclidean distance, which represents the simplest classifier permitting to understand the discriminative power of the proposed features. Classification accuracies have been estimated using the Leave One Out (LOO) scheme. In order to have a better insight into the proposed method, we compare the proposed approach with the individual separated audio and video processing; in particular, the 66 audio and the video FG patterns (see Sec. 3.4 and Sec. 3.3), extracted in the same time intervals of the AVC features (i.e. during the Potential Relation Interval - PRI), have been directly used as features to characterize the events. The classification accuracies for the three methods are presented in Table 1. Just at a first look of the table, one can notice a general benefit in integrating audio and visual information: audio-visual 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	Audio	Video	Audio-video
A	100.0%	86.3%	100.0%
B	60.8%	95.6%	95.6%
C	95.2%	85.7%	95.2%
D	62.1%	66.6%	89.3%

Table 1. LOO classification accuracies for the four different problems.

i) The scenario A is devoted to discriminate between making or receiving a phone call: clearly, most of the information is embedded into 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 gets worst results. It is important to notice that the audio-visual integration does not inhibit the information brought in the audio part.
ii) The scenarios B and C are characterized by two similar audio-visual situations. Regarding the audio part, there is a difference between talking in the lab or not talking, whereas regarding the video part there is the difference between switching on or off the lights. Actually, both single audio and video features get good results.
iii) The scenario D is the most complex and interesting. In this case, which involves 6 different classes, the integrated use of audio and video information permits to drastically improve the classification accuracy of about 25%. The tasks are complicated, and only a proper integration of audio and visual information could lead to a definite satisfactory classification results.

4.4. Clustering results

This last section reports results about clustering, in order to really discover patterns and natural groups of audio-video events. Given the automatically segmented dataset, we perform hierarchical clustering using the Ward scheme, and we consider the Euclidean distance as distance between elements. As in the classification task, we use a simple rule for performing clustering, in order to discover the expressivity of the AVD feature. We only set the number of clusters to 6, and let the algorithm to make the natural clusters. The resulting dendrogram is shown in Fig. 5, where in abscissas there are the situation labels. Observing the dendrogram, we can see that the underlying structure of the dataset is satisfactorily represented, but, obviously, there are some errors as the task is not easy. The most separated and well identified clusters are the situations 2 and 3: they are char-

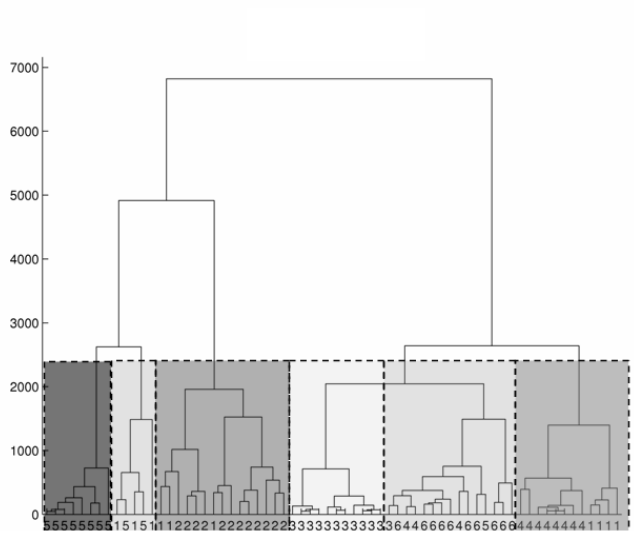

Figure 5. Hierarchical clustering dendrogram.

acterized by a clearly identifiable pattern (the ringing phone and the light on). Also cluster 5 is quite well identified; little confusion is present between patterns in cluster 4 and in cluster 6. This is a difficult case: the discrimination between them lies only in the fact that in the situation 4 the AV coupling happens among the opening door and the person entering in the scene (visual FG), together with 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 exited from the scene producing no visual FG.

The accuracy of clustering 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, also compared with single modalities results: clustering of audio patterns obtained an accuracy of 53.03%, whereas the video obtained 60.61%. Also in this case, it is evident the gain obtained by the fusion of audio-video information.

5. Conclusions

In this paper, a new method for characterizing audio visual events has been presented. Separate audio and video signals have been processed using two different adaptive modules, aimed at distilling 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 is realized by the means of the Audio-Video Concurrence matrix, a feature that permits to detect and segment AV events, as well as discriminating among them. Experimental results on real sequences have shown promising results, both in terms of classification and clustering. We are currently testing our approach with other baseline audio video classification and clustering approaches, in order to better validate the performances obtained.

References

[1] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Int. Conf. Computer Vision and Pattern Recognition (CVPR '99)*, vol. 2, 1999, pp. 246–252. 1, 2, 3

[2] M. Cristani, M. Bicego, and V. Murino, "On-line adaptive background modelling for audio surveillance," in *Proceedings of International Conference on Pattern Recognition (ICPR 2004)*, 2004, pp. 399–402. 1

[3] T. Darrell, J. Fisher, and K. Wilson, "Geometric and statistical approaches to audiovisual segmentation for unthetered interaction," CLASS Project, Tech. Rep., 2002. 1

[4] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. John Wiley and Sons, 2001. 1

[5] N. Checka and K. Wilson, "Person tracking using audio-video sensor fusion," MIT Artificial Intelligence Laboratory, Tech. Rep., 2002. 2

[6] X. Zou and B. Bhanu, "Tracking humans using multimodal fusion," in *CVPR '05: Proceedings of the 2005 Conference on Computer Vision and Pattern Recognition (CVPR '05)*, 2005. 2

[7] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Proc. Neural Information Processing (NIPS 2000)*, 2000. 2

[8] J. Kay, "Feature discovery under contextual supervision using mutual information," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 4. IEEE Computer Society, 1992, pp. 79–84. 2

[9] D. Zhang, D. Gatica-Perez, and S. Bengio, "Semi-supervised adapted HMMs for unusual event detection," in *CVPR '05: Proceedings of the 2005 Conference on Computer Vision and Pattern Recognition (CVPR '05)*, 2005, pp. 611–618. 2

[10] M. Petkovic and W. Jonker, *Content-Based Video Retrieval : A Database Perspective (Multimedia Systems and Applications)*. Springer, 2003. 2