Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2011, Article ID 485738, 11 pages
doi:10.1155/2011/485738

Research Article

Acoustic Event Detection Based on Feature-Level Fusion of Audio and Video Modalities

Taras Butko, Cristian Canton-Ferrer, Carlos Segura, Xavier Giró, Climent Nadeu, Javier Hernando, and Josep R. Casas

Department of Signal Theory and Communications, TALP Research Center, Technical University of Catalonia, Campus Nord, Ed. D5, Jordi Girona 1-3, 08034 Barcelona, Spain

Correspondence should be addressed to Taras Butko, taras.butko@upc.edu

Received 20 May 2010; Revised 30 November 2010; Accepted 14 January 2011

Academic Editor: Sangjin Hong

Acoustic event detection (AED) aims at determining the identity of sounds and their temporal position in audio signals. When applied to spontaneously generated acoustic events, AED based only on audio information shows a large amount of errors, which are mostly due to temporal overlaps. Actually, temporal overlaps accounted for more than 70% of errors in the real-world interactive seminar recordings used in CLEAR 2007 evaluations. In this paper, we improve the recognition rate of acoustic events using information from both audio and video modalities. First, the acoustic data are processed to obtain both a set of spectrotemporal features and the 3D localization coordinates of the sound source. Second, a number of features are extracted from video recordings by means of object detection, motion analysis, and multicamera person tracking to represent the visual counterpart of several acoustic events. A feature-level fusion strategy is used, and a parallel structure of binary HMM-based detectors is employed in our work. The experimental results show that information from both the microphone array and video cameras is useful to improve the detection rate of isolated as well as spontaneously generated acoustic events.

1. Introduction

The detection of the acoustic events (AEs) naturally produced in a meeting room may help to describe the human and social activity. The automatic description of interactions between humans and environment can be useful for providing implicit assistance to the people inside the room, providing context-aware and content-aware information requiring a minimum of human attention or interruptions [1], providing support for high-level analysis of the underlying acoustic scene, and so forth. In fact, human activity is reflected in a rich variety of AEs, either produced by the human body or by objects handled by humans. Although speech is usually the most informative AE, other kind of sounds may carry useful cues for scene understanding. For instance, in a meeting/lecture context, we may associate a chair moving or door noise to its start or end, cup clinking to a coffee break, or footsteps to somebody entering or leaving. Furthermore, some of these AEs are tightly coupled with human behaviors or psychological states: paper wrapping may denote tension; laughing, cheerfulness; yawning in the middle of a lecture, boredom; keyboard typing, distraction from the main activity in a meeting; clapping during a speech, approval. Acoustic event detection (AED) is also useful in applications as multimedia information retrieval, automatic tagging in audio indexing, and audio context classification. Moreover, it can contribute to improve the performance and robustness of speech technologies such as speech and speaker recognition and speech enhancement.

Detection of acoustic events has been recently performed in several environments like hospitals [2], kitchen rooms [3], or bathrooms [4]. For meeting-room environments, the task of AED is relatively new; however, it has already been evaluated in the framework of two international evaluation campaigns: in CLEAR (Classification of Events, Activities, and Relationships evaluation campaigns) 2006 [5], by three participants, and in CLEAR 2007 [6], by six participants. In the last evaluations, 5 out of 6 submitted systems showed accuracies below 25%, and the best system got 33.6% accuracy [7]. In most submitted systems, the standard

combination of cepstral coefficients and hidden Markov model (HMM) classifiers widely used in speech recognition is exploited. It has been found that the overlapping segments account for more than 70% of errors produced by every submitted system.

The overlap problem may be tackled by developing more efficient algorithms either at the signal level using source separation techniques like independent component analysis [8]; at feature level, by means of using specific features [9] or at the model level [10]. Another approach is to use an additional modality that is less sensitive to the overlap phenomena present in the audio signal. In fact, most of human-produced AEs have a visual correlate that can be exploited to enhance the detection rate. This idea was first presented in [11], where the detection of footsteps was improved by exploiting the velocity information obtained from a video-based person-tracking system. Further improvement was shown in our previous papers [12, 13], where the concept of multimodal AED is extended to detect and recognize the set of 11 AEs. In that work, not only video information but also acoustic source localization information was considered.

In the work reported here, we use a feature-level fusion strategy and a structure of the HMM-based system which considers each class separately, using a one-against-all strategy for training. To deal with the problem of insufficient number of AE occurrences in the database we used so far, 1 additional hour of training material has been recorded for the presented experiments. Moreover, video feature extraction is extended to 5 AE classes, and the additional "Speech" class is also evaluated in the final results. A statistical significance test is performed individually for each acoustic event. The main contribution of the presented work is twofold. First, the use of video features, which are new for the meeting-room AED task. Since the video modality is not affected by acoustic noise, the proposed features may improve AED in spontaneous scenario recordings. Second, the inclusion of acoustic localization features, which, in combination with usual spectrotemporal audio features, yield further improvements in recognition rate.

The rest of this paper is organized as follows. Section 2 describes the database and metrics used to evaluate the performance. The feature extraction process from audio and video signals is described in Sections 3 and 4, respectively. In Section 5, both the detection system and the fusion of different modalities are described. Section 6 presents the obtained experimental results, and, finally, Section 7 provides some conclusions.

2. Database and Metrics

There are several publicly available multimodal databases designed to recognize events, activities, and their relationships in interaction scenarios [1]. However, these data are not well suited to audiovisual AED since the employed cameras do not provide a close view of the subjects under study. A new database has been recorded with 5 calibrated cameras at a resolution of 768×576 at 25 fps, and 6 T-shaped 4-microphone clusters are also employed, sampling the acoustic signal at 44.1 kHz. Synchronization among all sensors is fulfilled. This database includes two kinds of datasets: 8 recorded sessions of isolated AEs, where 6 different participants performed 10 times each AE, and a spontaneously generated dataset which consists of 9 scenes about 5 minutes long with 2 participants that interact with each other in a natural way, discuss certain subject, drink coffee, speak on the mobile phone, and so forth. Although the interactive scenes were recorded according to a previously elaborated scenario, we call this type of recordings "spontaneous" since the AEs were produced in a realistic seminar style with possible overlap with speech. Besides, all AEs appear with a natural frequency; for instance, applause appears much less frequently (1 instance per scene) than chair moving (around 8–20 instances per scene). Manual annotation of the data has been done to get an objective performance evaluation. This database is publicly available from the authors.

The considered AEs are presented in Table 1, along with their number of occurrences.

The metric referred to AED-ACC (1) is employed to assess the final accuracy of the presented algorithms. This metric is defined as the F-score (the harmonic mean between precision and recall)

$$\text{AED-ACC} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{1}$$

where

$$\begin{aligned} \text{Precision} &= \frac{\text{number of correct system output AEs}}{\text{number of all system output AEs}}, \\ \text{Recall} &= \frac{\text{number of correctly detected reference AEs}}{\text{number of all reference AEs}}. \end{aligned} \tag{2}$$

A system output AE is considered correct if at least one of two conditions is met. (1) There exists at least one reference AE whose temporal centre is situated between the timestamps of the system output AE, and the labels of the system output AE and the reference AE are the same. (2) Its temporal centre lies between the timestamps of at least one reference AE, and the labels of both the system output AE and the reference AE are the same. Similarly, a reference AE is considered correctly detected if at least one of two conditions is met. (1) There exists at least one system output AE whose temporal centre is situated between the timestamps of the reference AE, and the labels of both the system output AE and the reference AE are the same. (2) Its temporal centre lies between the timestamps of at least one system output AE, and the labels of the system output AE and the reference AE are the same.

The AED-ACC metric was used in the last CLEAR 2007 [6] international evaluation, supported by the European Integrated project CHIL [1] and the US National Institute of Standards and Technology (NIST).

3. Audio Feature Extraction

The basic features for AED come from the audio signals. In our work, a single audio channel is used to compute a set of

Table 1: Number of occurrences per acoustic event class for train and test data.

Acoustic event	Label	Number of occurrences	
		Isolated	Spontaneously generated
Door knock	[kn]	79	27
Door open/slam	[ds]	256	82
Steps	[st]	206	153
Chair moving	[cm]	245	183
Spoon/cup jingle	[cl]	96	48
Paper work	[pw]	91	146
Key jingle	[kj]	82	41
Keyboard typing	[kt]	89	81
Phone ring	[pr]	101	29
Applause	[ap]	83	9
Cough	[co]	90	24
Speech	[sp]	74	255

audio spectrotemporal (AST) features. That kind of features, which are routinely used in audio and speech recognition [2–4, 7, 10, 14], describe the spectral envelope of the audio signal within a frame and its temporal evolution along several frames. However, this type of information is not sufficient to deal with the problem of AED in presence of temporal overlaps. In the work reported here, we firstly propose to use the additional audio information from a microphone array available in the room, by extracting features which describe the spatial location of the produced AE in the 3D space. Although both types of features (AST and localization features) are originated from the same physical acoustic source, they are regarded here as features belonging to two different modalities.

3.1. Spectrotemporal Audio Features. A set of audio spectrotemporal features is extracted to describe every audio signal frame. In our experiments, the frame length is 30 ms with 20 ms shift, and a Hamming window is applied. There exist several alternative ways of parametrically representing the spectrum envelope of audio signals. The mel-cepstrum representation is the most widely used in recognition tasks. In our work, we employ a variant called frequency-filtered (FF) log filter-bank energies (LFBEs) [14]. It consists of applying, for every frame, a short-length FIR filter to the vector of log filter-bank energies vector, along the frequency variable. The transfer function of the filter is $z\text{-}z^{-1}$, and the end points are taken into account. That type of features has been successfully applied not only to speech recognition but also to other speech technologies like speaker recognition [15]. In the experiments, 16 FF-LFBEs are used, along with their first temporal derivatives, the latter representing the temporal evolution of the envelope. Therefore, a 32-dimensional feature vector is used.

3.2. Localization Features. In order to enhance the recognition results, acoustic localization features are used in combination with the previously described AST features. In our case, as the characteristics of the room are known beforehand (Figure 1(a)), the position (x, y, z) of the acoustic source may carry useful information. Indeed, some acoustic events can only occur at particular locations, like door slam and door knock can only appear near the door, or footsteps and chair moving events take place near the floor. Based on this fact, we define a set of metaclasses that depend on the position where the acoustic event is detected. The proposed metaclasses and their associated spatial features are "near door" and "far door," related to the distance of the acoustic source to the door, and "below table," "on table," and "above table" metaclasses depending on the z-coordinate of the detected AE. The height-related metaclasses are depicted in Figure 1(b), and their likelihood function modelled via Gaussian mixture models (GMMs) can be observed in Figure 2(b). It is worth noting that the z-coordinate is not a discriminative feature for those AEs that are produced at the similar height.

The acoustic localization system used in this work is based on the SRP-PHAT [16] localization method, which is known to perform robustly in most scenarios. The SRP-PHAT algorithm is briefly described in the following. Consider a scenario provided with a set of N_M microphones from which we choose a set of microphone pairs, denoted as Ψ. Let X_i and X_j be the 3D location of two microphones i and j. The time delay of a hypothetical acoustic source placed at $x \in R^3$ is expressed as

$$\tau_{x,i,j} = \frac{\|x - x_i\| - \left\|x - x_j\right\|}{s}, \tag{3}$$

where s is the speed of sound. The 3D space to be analyzed is quantized into a set of positions with typical separations of 5 to 10 cm. The theoretical TDoA $\tau_{x,i,j}$ from each exploration position to each microphone pair is precalculated and stored. PHAT-weighted cross correlations of each microphone pair are estimated for each analysis frame [17]. They can be expressed in terms of the inverse Fourier transform of the estimated cross-power spectral density $G_{i,j}(f)$ as follows:

$$R_{i,j}(\tau) = \int_{-\infty}^{\infty} \frac{G_{i,j}(f)}{\left|G_{i,j}(f)\right|} e^{j2\pi f\tau} df. \tag{4}$$

The contribution of the cross correlation of every microphone pair is accumulated for each exploration region using the delays precomputed in (4). In this way, we obtain an acoustic map at every time instant, as depicted in Figure 2(a). Finally, the estimated location of the acoustic source is the position of the quantized space that maximizes the contribution of the cross correlation of all microphone pairs

$$\hat{x} = \underset{x}{\operatorname{argmax}} \sum_{i,j\in\Psi} R_{i,j}\left(\tau_{x,i,j}\right). \tag{5}$$

The sum of the contributions of each microphone pair crosscorrelation gives a value of confidence of the estimated position, which is assumed to be well correlated with the likelihood of the estimation.

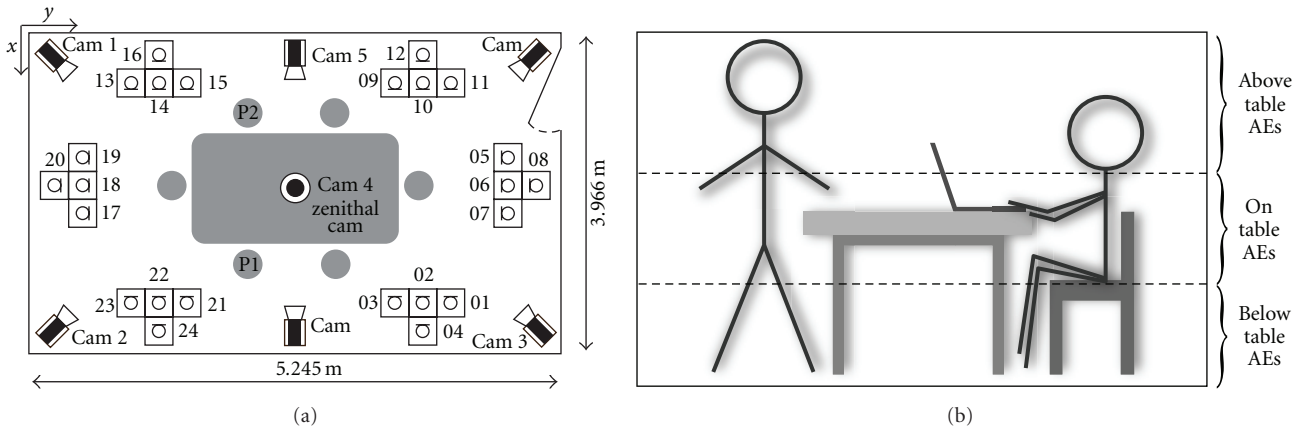


FIGURE 1: (a) The top view of the room. (b) The three categories along the vertical axis.

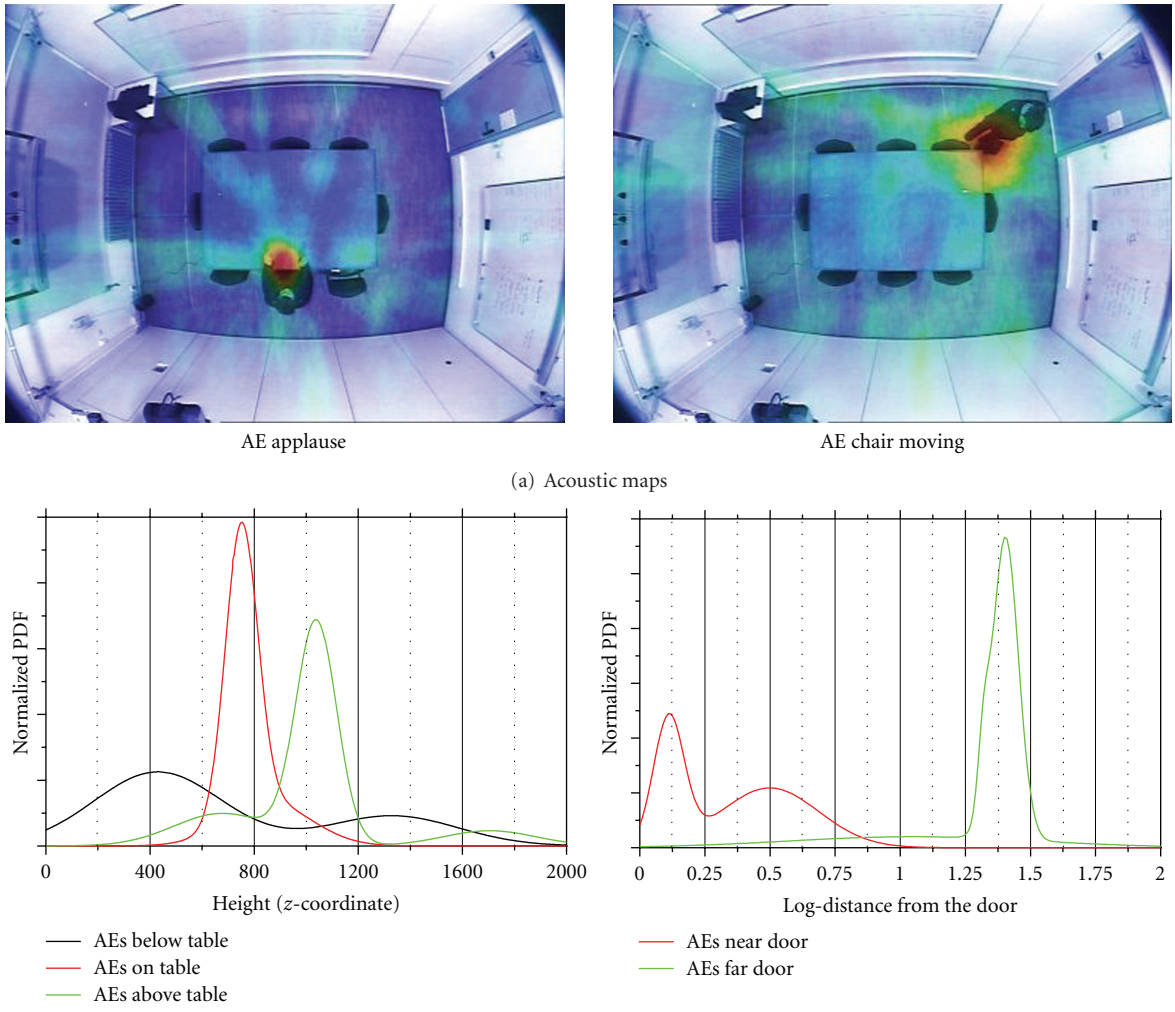


FIGURE 2: Acoustic localization. In (a), acoustic maps corresponding to two AEs overlaid to a zenithal camera view of the analyzed scenario. In (b), the likelihood functions modelled by GMMs.

4. Video Feature Extraction

AED is usually addressed from an audio perspective only. Typically, low acoustic energy AEs as paper wrapping, keyboard typing, or footsteps are hard to be detected using only the audio modality. The problem becomes even more challenging in the case of signal overlaps. Since the human-produced AEs have a visual correlate, it can be exploited to enhance the detection rate of certain AEs. Therefore, a number of features are extracted from video recordings by means of object detection, motion analysis, and multicamera person tracking to represent the visual counterpart of 5 classes of AEs. From the audio perspective, the video modality has an attractive property; the disturbing acoustic noise usually does not have a correlate in the video signal. In this section, several video technologies which provide useful features for our AED task are presented.

4.1. Person Tracking Features. Tracking of multiple people present in the analysis area basically produces two figures associated with each target position and velocity. As it has been commented previously, acoustic localization is directly associated with some AEs but, for the target's position obtained from video, this assumption cannot be made. Nonetheless, target's velocity is straightforward associated with footstep AE. Once the position of the target is known, an additional feature associated with the person can be extracted: height. When analyzing the temporal evolution of this feature, sudden changes of it are usually correlated with chair moving AE, that is, when the person sits down or stands up. The derivative of height position along the time is employed to address the "Chair moving" detection. Multiple cameras are employed to perform tracking of multiple interacting people in the scene, applying the real-time performance algorithm presented in [18]. This technique exploits spatial redundancy among camera views towards avoiding occlusion and perspective issues by means of a 3D reconstruction of the scene. Afterwards, an efficient Monte Carlo-based tracking strategy retrieves an accurate estimation of both the location and velocity of each target at every time instant. An example of the performance of this algorithm is shown in Figure 3(a). The likelihood functions of velocity feature for class "Steps" and metaclass "Nonsteps" are shown in Figure 3(b).

4.2. Color-Specific MHE Features. Some AEs are associated with motion of objects around the person. In particular, we would like to detect a motion of a white object in the scene that can be associated to paper wrapping (under the assumption that a paper sheet is distinguishable from the background color). In order to address the detection of white paper motion, a close-up camera focused on the front of the person under study is employed. Motion descriptors introduced by [19], namely, the motion history energy (MHE) and image (MHI), have been found useful to describe and recognize actions. However, in our work, only the MHE feature is exploited, since the MHI descriptor encodes the structure of the motion, that is, how the action is executed; this cue does not provide any useful information to increase the classifier performance. Every pixel in the MHE image contains a binary value denoting whether motion has occurred in the last τ frames at that location. In the original technique, silhouettes were employed as the input to generate these descriptors, but they are not appropriate in our context since motion typically occurs within the silhouette of the person. Instead, we propose to generate the MHE from the output of a pixel-wise color detector, hence performing a color/region-specific motion analysis that allows distinguishing motion for objects of a specific color. For paper motion, a statistic classifier based on a Gaussian model in RGB is used to select the pixels with whitish color. In our experiments, $\tau = 12$ frames produced satisfactory results. Finally, a connected component analysis is applied to the MHE images, and some features are computed over the retrieved components (blobs). In particular, the area of each blob allows discarding spurious motion. In the paper motion case, the size of the biggest blob in the scene is employed to address paper wrapping AE detection. An example of this technique is depicted in Figure 4.

4.3. Object Detection. Detection of certain objects in the scene can be beneficial to detect some AEs such as phone ringing, cup clinking, or keyboard typing. Unfortunately, phones and cups are too small to be efficiently detected in our scenario, but the case of a laptop can be correctly addressed. In our case, the detection of laptops is performed from a zenithal camera located at the ceiling. The algorithm initially detects the laptop's screen and keyboard separately and, in a second stage, assesses their relative position and size. Captured images are segmented to create an initial partition of 256 regions based on color similarity. These regions are iteratively fused to generate a binary partition tree (BPT), a region-based representation of the image that provides segmentation at multiple scales [20]. Starting from the initial partition, the BPT is built by iteratively merging the two most similar and neighboring regions, defining a tree structure whose leaves represent the regions at the initial partition and the root corresponds to the whole image (see Figure 5(a)). Thanks to this technique, the laptop parts may be detected not only at the regions in the initial partition but also at some combinations of them, represented by the BPT nodes. Once the BPT is built, visual descriptors are computed for each region represented at its nodes. These descriptors represent color, area, and location features of each segment.

The detection problem is posed as a traditional pattern recognition case, where a GMM-based classifier is trained for the screen and keyboard parts. A subset of ten images representing the laptop at different positions in the table has been used to train a model based on the region-based descriptors of each laptop part, as well as their relative position and sizes. An example of the performance of this algorithm is shown in Figure 5(b). For further details on the algorithm, the reader is referred to [21].

4.4. Door Activity Features. In order to visually detect door slam AE, we considered exploiting the a priori knowledge

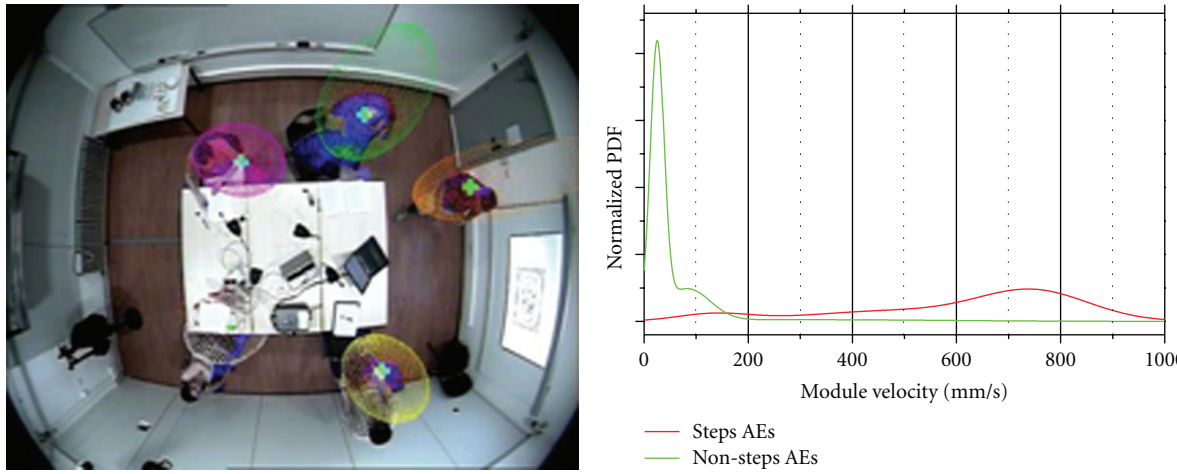


Figure 3: Person tracking. In (a), the output of the employed algorithm in a scenario involving multiple targets. In (b), the likelihood functions of the velocity feature corresponding to "Steps" and "Nonsteps" AEs.

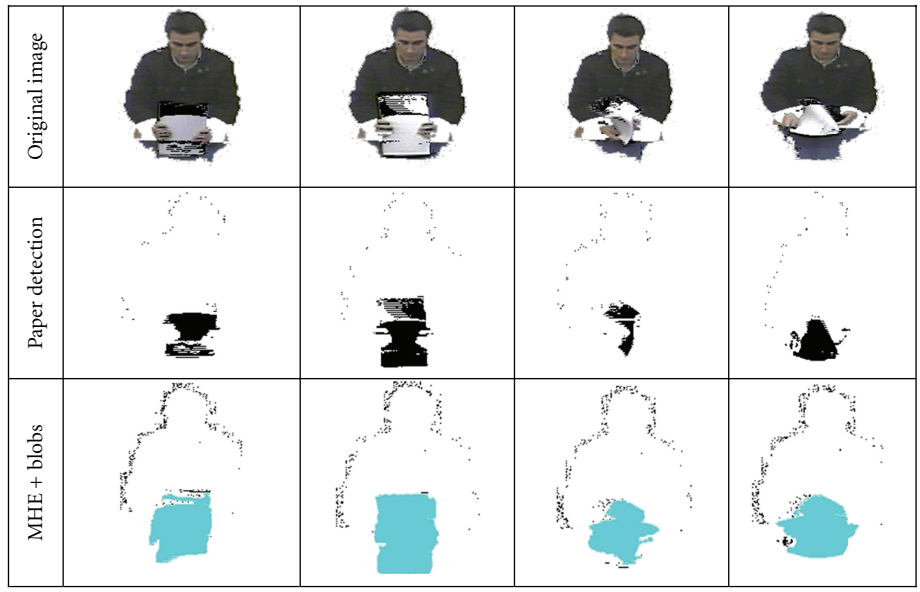


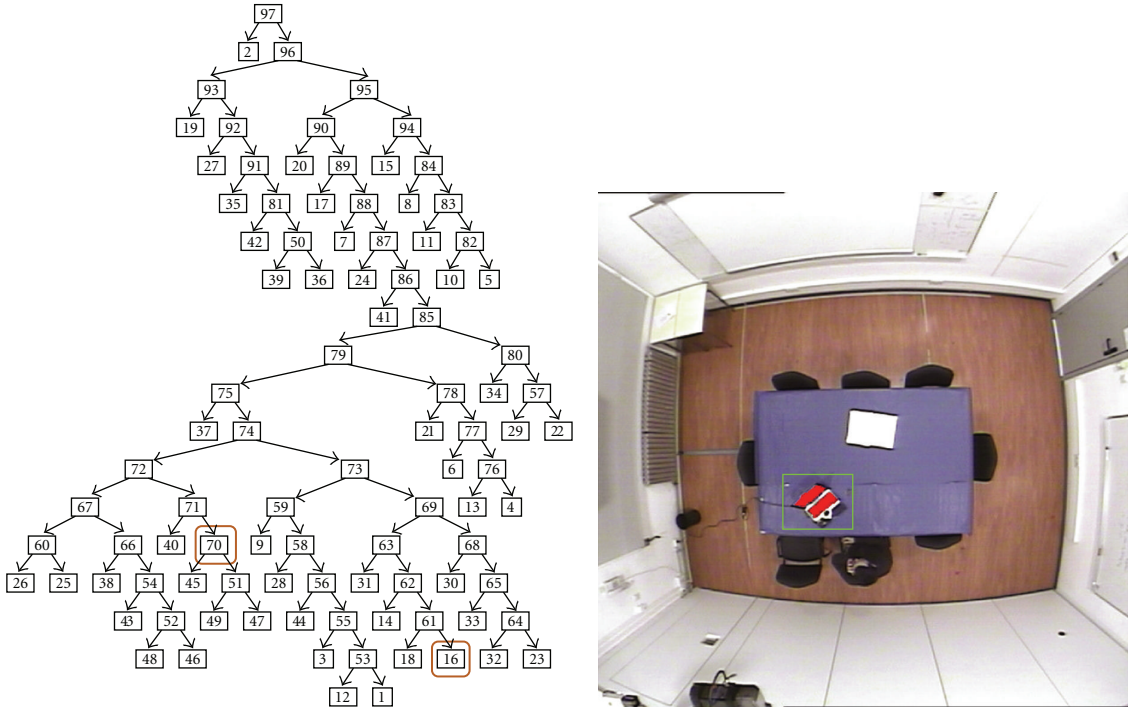
Figure 4: Paper wrapping feature extraction.

about the physical location of the door. Analyzing the zenithal camera view, activity near the door can be addressed by means of a foreground/background pixel classification [22]. The amount of foreground pixels in the door area will indicate that a person has entered or exited, hence allowing a visual detection of door slam AE.

5. Multimodal Acoustic Event Detection

Once the informative features related to the AEs of interest are extracted for every input modality, a multimodal-based classification is performed. The overall diagram of the proposed system is depicted in Figure 6. Three data sources are combined together: two come from audio and one from video. The first is obtained from single channel audio processing and consists of AST features. The second is obtained from microphone array processing and consists of the 3D location of the audio source. And the third is obtained from multiple cameras covering the scenario and consists of video-based features related to several AEs. The three types of features are concatenated together (feature-level fusion) and supplied to the corresponding binary detector from the set of 12 detectors that work in parallel.

5.1. Binary Detection System. In the work reported here, each AE class is modeled via hidden Markov model (HMM) with GMM observation probability distributions, like in [13], and the Viterbi decoding algorithm is used for segmentation.



(a) (b)

Figure 5: Object detection. In (a), the binary tree representing the whole image as a hierarchy. Regions corresponding to the screen and keyboard regions are identified within the tree. In (b), the detection of a laptop from zenithal view.

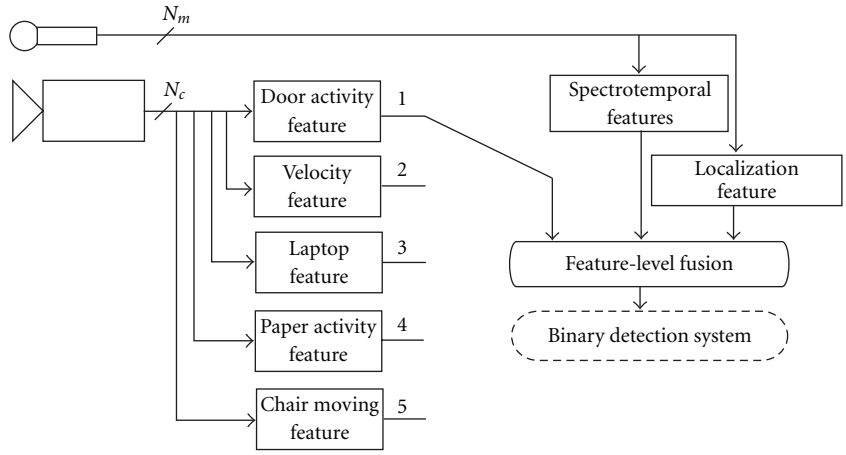


Figure 6: System flowchart.

Although the multiclass segmentation is usually performed within a single pass, in our work, we exploit the parallel structure of the binary detectors depicted in Figure 7. Firstly, the input signal is processed by each binary detector independently (the total number of detectors is equal to the number of AE classes), thus segmenting the input signal in intervals either as "Class" or "Nonclass." Using the training approach known as one-against-all method [23], all the classes different from "Class" are used to train the "Nonclass" model. The models for "Class" and "Nonclass" are HMMs with 3 emitting states and left-to-right connected state transitions. The observation distributions of the states are Gaussian mixtures with continuous densities and consist of 5 components with diagonal covariance matrices. Secondly, the sequences of decisions from each binary detector are combined together to get the final decision.

The proposed architecture with 12 separate HMM-based binary detectors working in parallel has several advantages.

(1) For each particular AE, the best set of features is used. The features which are useful for detecting one class are not necessarily useful for other classes. In our case, the video features are used only for detecting some particular classes.

Table 2: Monomodal recognition results.

	AST (%)	Video (%)	Localization (%)
Door knock	97.20	—	82.95
Door slam	93.95	79.96	
Chair moving	94.73	77.28	83.15
Steps	60.94	75.60	
Paper work	94.10	91.42	86.31
Keyboard	95.57	81.98	
Cup clink	95.47	—	
Key jingle	89.73	—	67.70
Phone ring	89.97	—	
Applause	93.24	—	
Cough	93.19	—	
Speech	86.25	—	

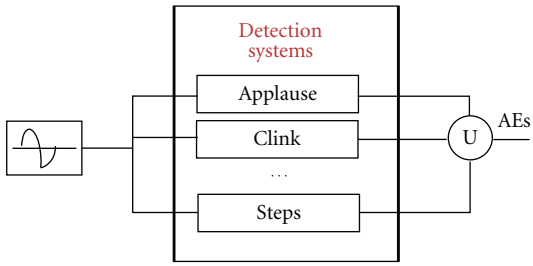


Figure 7: A set of binary detectors working in parallel.

(2) The tradeoff between the number of misses and false alarms can be optimized for each particular AE class.

(3) In the case of overlapped AEs, the proposed system can provide multiple decisions for the same audio segment.

However, this architecture requires N binary detectors, where N is the total number of AE classes. This makes the detection process more complex in the case of a large number of AE classes. In [13], it was shown that the detection system based on the set of binary detectors working in parallel shows higher accuracy rate than the AED system based on one multiclass detector.

5.2. Fusion of Different Modalities. The information fusion can be done on data, feature, and decision levels. Data fusion is rarely found in multimodal systems because raw data is usually not compatible among modalities. For instance, audio is represented by one-dimensional vector of samples, whereas video is organized in two-dimensional frames. Concatenating feature vectors from different modalities into one super vector is a possible way for combining audio and visual information. This approach has been reported, for instance, in [24], for multimodal speech recognition.

5.2.1. Feature-Level Fusion Approach. In this work, we use an HMM-GMM approach with feature-level fusion, which is implemented by concatenating the feature sets X_1, X_2, and X_3 from 3 different modalities in one super-vector $Z = [X_1\ X_2\ X_3]$. In our framework, X_1 corresponds to 32 AST features; X_2 corresponds to 1 localization feature (either z-position or distance from the door); X_3 corresponds to 1 video-based feature (see Figure 6). In total, a 34-dimensional feature vector is obtained for those 5 classes of AEs for which the video modality is taken into account ("door slam", "steps", "keyboard typing", "paper wrapping," and "chair moving"). For the rest of AEs, only X_1 and X_2 are used (in this case the feature vector has 33 components).

Then, the likelihood of that observation super vector at state j and time t is calculated every frame of 20 ms as

$$b_Z(t) = \sum_m p_m N(Z_t; \mu_m; \Sigma_m), \tag{6}$$

where $N(\cdot; \mu; \Sigma)$ is a multivariate Gaussian pdf with mean vector μ and covariance matrix Σ, and p_m are the mixture weights. Assuming uncorrelated feature streams, diagonal covariance matrices are considered.

5.2.2. Dealing with Missing Features. The feature-level fusion becomes difficult task when some features are missing. Although the AST features can be extracted at every time instance, the feature that corresponds to the localization of acoustic source has undefined value in the absence of any acoustic activity. The same situation happens with the position in 3D space of the person while nobody is inside the room. There are two major approaches to solve this problem [25].

(a) Feature-vector imputation: estimate the missed feature components to reconstruct a complete feature vector and use it for recognition.

(b) Classifier modification: modify the classifier to perform recognition using existing features (the most usual method is marginalization).

In fact, both of the above-mentioned cases of missing features are associated with the silence AE. This way the fact that the feature is missing may carry useful information about underlying acoustic scene. So, we impute the missing features (x, y, z coordinates) with the predefined "synthetic" value (we use -1 value in our experiments). In this case, we explicitly assign the 3D "position" of silence event to have $(-1,\ -1,\ -1)$ value.

6. Experiments

In order to assess the performance of the proposed multimodal AED system and show the advantages of the proposed feature sets, the database of isolated AEs described in Section 2 was used for both training and testing: 8 sessions were randomly permuted; odd index numbers were assigned to training and even index numbers to testing. Six permutations were used in the experiments. The subset of spontaneously generated AEs was used in the final experiments in order to check the adequateness of the multimodal fusion with real world data.

Table 3: Confusion matrix corresponding to the baseline system (the results are presented in %).

	kn	ds	cm	st	pw	kt	cl	kj	pr	ap	co	sp
kn	98.8	0.4	0	0	0	0	0	0	0	0	0.8	0
ds	0.3	82.0	0	14.8	0.1	1.2	0.4	0.1	0.2	0.2	0.2	0
cm	0.9	0.4	93.8	4.0	0.4	0	0	0	0	0	0.1	0.3
st	0	18.1	13.8	65.4	1.2	0.5	0	0	0.2	0.4	0	0.4
pw	0	0.3	0	0.3	85.6	10.5	0	1.0	0.3	2.0	0	0
kt	0	0	0	0	0	98.9	0	0.8	0.4	0	0	0
cl	0	2.0	0	0	0	0	94.9	1.0	2.0	0	0	0
kj	0	0	0	0	5.0	0.8	0	89.5	4.7	0	0	0
pr	0	0	0	0	0	0	0	1.0	87.8	0.3	0	10.9
ap	0	0	0	0	1.2	0	1.2	0	0	97.6	0	0
co	6.9	0.4	0	0	0	0	0	0	0	0	92.4	0.4
sp	1.8	0.7	0	5.8	0	0	0	0	3.6	0	7.6	80.6

Table 4: Fusion of different modalities using isolated and spontaneously generated AEs.

AEs	Isolated					Spontaneously generated			
	AST	AST+L	AST+V	AST+L+V	*P*-value	AST	AST+L	AST+V	AST+L+V
Door knock	97.20	98.81	97.20	98.81	.05	88.72	90.45	88.72	90.45
Door slam	93.95	95.35	97.06	96.72	.01	75.45	82.89	85.04	87.36
Chair moving	94.73	95.18	95.24	95.93	.09	83.89	84.32	84.12	84.82
Steps	60.94	72.51	78.09	77.25	.04	58.56	57.12	67.12	66.58
Paper work	94.10	94.19	95.16	95.07	.30	65.14	62.61	73.18	79.32
Keyboard	95.57	95.96	96.56	96.72	.37	71.69	78.37	79.68	80.50
Cup clink	95.47	94.03	95.47	94.03	.86	90.35	86.08	90.35	86.08
Key jingle	89.73	88.00	89.73	89.60	.52	52.09	44.12	52.09	44.12
Phone	89.97	88.09	89.97	88.79	.64	87.98	90.45	87.98	90.45
Applause	93.24	94.91	93.24	94.91	.13	84.06	84.65	84.06	84.65
Cough	93.19	94.20	93.19	94.20	.35	76.47	82.36	76.47	82.36
Speech	86.25	85.47	86.25	85.47	.62	83.66	83.12	83.66	83.12
Average	90.36	91.39	92.26	92.29	—	76.51	77.21	79.37	79.98

The detection results for each monomodal detection system are presented in Table 2 (for the database of isolated AEs only). The baseline system (first column) is trained with the 32 spectrotemporal features, while the other two systems use only one feature coming from either the video or the localization modality, respectively. As we see from the table, the baseline detection system shows high recognition rates for almost all AEs except the class "Steps" that is much better detected with the video-based AED system. The recognition results for the video-based system are presented only for those AEs for which video counterpart is taken into consideration. In the case of localization-based AED system, the results are presented only for each category rather than the particular AE class. In fact, using the localization information, we are able to detect just the category but not the AE within it.

The confusion matrix that corresponds to the baseline detection system is presented in Table 3, which presents the percentage of hypothesized AEs (rows) that are associated to the reference AEs (columns), so that all the numbers out of the main diagonal correspond to confusions. This table shows that some improvement may be achieved by adding localization-based features. For instance, although the "below-table" AEs ("Chair moving" and "Steps") are mainly confused with each other, there is still some confusion among these two AEs and the AEs from other categories.

The final detection results for isolated and spontaneously generated AEs are presented in Table 4. The first column corresponds to the baseline system (that uses the 32-dimensional AST feature vector). The next columns correspond to the fusion of baseline features with the localization feature, the video feature, and the combination of both of them, respectively. The last column shows the P value of the statistical significance of the AST+L+V test in relation to the baseline system. If P_1 and P_2 are the accuracy measures for the baseline and the multimodal AED system, respectively, the null hypothesis H_0 is $P_1 \geq P_2$; and the alternative hypothesis H_1 is $P_1 < P_2$. Assuming a standard level of significance at 95%, a P value that is less than .05 implies the rejection of the null hypothesis or, in other words, it means that the result is statistically significant.

Although the AST+L+V system improves the baseline system for most of the isolated AEs, a statistically significant improvement is only obtained for the classes "Door

slam", "Door knock", and "Steps." For the data subset of spontaneously generated AEs, a significant improvement in the detection of some low energy AEs ("Steps", "Paper work", "Keyboard typing") is achieved. The best relative improvement corresponds to the "Steps" class. Other AEs have slightly improved their detection rates. In average, 15% relative error-rate reduction for isolated AEs and 21% for spontaneously generated AEs are achieved.

As it can be observed, the video information improves the baseline results for the five classes for which video information is used, especially in the case of spontaneously generated AEs where the acoustic overlaps happen more frequently. Therefore, the recognition rate of those classes considered as "difficult" (usually affected by overlap or of low energy) increases.

Acoustic localization features improve the recognition accuracy for some AEs, but for other events, it is decreased. One of the reasons of such behavior is the mismatch between training and testing data for spontaneously generated AEs. For instance, the "Cup clink" AE in spontaneous conditions often appears when the person is standing, which is not the case for isolated AEs. Another reason is that, for overlapped AEs, the AE with higher energy will be properly localized while the other overlapped AE will be masked. Additionally, according to the confusion matrix (Table 3), the main confusion among AEs happens inside the same category, so that the audio localization information is not able to contribute significantly.

7. Conclusions and Future Work

In this paper, a multimodal system based on a feature-level fusion approach and a one-against-all detection strategy has been presented and tested with a new audiovisual database. The acoustic data is processed to obtain a set of spectrotemporal features and the localization coordinates of the sound source. Additionally, a number of features are extracted from the video signals by means of object detection, motion analysis, and multicamera person tracking to represent the visual counterpart of several AEs. Experimental results show that information from the microphone array as well as the video cameras facilitates the task of AED for both datasets of AEs: isolated and spontaneously generated. Since the video signals are not affected by acoustic noise, a significant error-rate reduction is achieved due to the video modality. The acoustic localization features also improve the results for some particular classes of AEs. The combination of all features produced higher recognition rates for most of the classes, being the improvement statistically significant for a few of them.

Future work will be devoted to extend the multimodal AED system to other classes as well as the elaboration of new multimodal features and fusion techniques.

Acknowledgments

This work has been funded by the Spanish Project SAPIRE (no. TEC2007-65470). T. Butko is partially supported by a grant from the Catalan autonomous government.

References

[1] A. Waibel and R. Stiefelhagen, *Computers in the Human Interaction Loop*, Springer, New York, NY, USA, 2009.

[2] M. Vacher, D. Istrate, L. Besacier, E. Castelli, and J. Serignat, "Smart audio sensor for telemedicina," in *Proceedings of the Smart Object Conference*, 2003.

[3] M. Stäger, P. Lukowicz, N. Perera, T. Von Büren, G. Tröster, and T. Starner, "Sound Button: design of a low power wearable audio classification system," in *Proceedings of the International Symposium on Wearable Computers (ISWC '03)*, pp. 12–17, 2003.

[4] C. Jianfeng, Z. Jianmin, A. H. Kam, and L. Shue, "An automatic acoustic bathroom monitoring system," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '05)*, pp. 1750–1753, May 2005.

[5] CLEAR, "Classification of Events, Activities and Relationships. Evaluation and Workshop," 2006, http://isl.ira.uka.de/clear06.

[6] CLEAR, "Classification of Events, Activities and Relationships. Evaluation and Workshop," 2007, http://www.clear-evaluation.org.

[7] A. Temko, C. Nadeu, and J.-I. Biel, "Acoustic event detection: SVM-based system and evaluation setup in CLEAR," in *Multimodal Technologies for Perception of Humans*, vol. 4625 of *LNCS*, pp. 354–363, Springer, New York, NY, USA, 2008.

[8] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, New York, NY, USA, 2001.

[9] S. N. Wrigley, G. J. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 84–91, 2005.

[10] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, 2009.

[11] T. Butko, A. Temko, C. Nadeu, and C. Canton, "Fusion of audio and video modalities for detection of acoustic events," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH '08)*, pp. 123–126, 2008.

[12] C. Canton-Ferrer, T. Butko, C. Segura et al., "Audiovisual event detection towards scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 81–88, June 2009.

[13] T. Butko, C. Canton-Ferrer, C. Segura et al., "Improving detection of acoustic events using audiovisual data and feature level fusion," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH '09)*, pp. 1147–1150, September 2009.

[14] C. Nadeu, D. Macho, and J. Hernando, "Time and frequency filtering of filter-bank energies for robust HMM speech recognition," *Speech Communication*, vol. 34, no. 1-2, pp. 93–114, 2001.

[15] J. Luque and J. Hernando, "Robust speaker identification for meetings: UPC CLEAR-07 meeting room evaluation system," in *Multimodal Technologies for Perception of Humans*, vol. 4625 of *LNCS*, pp. 266–275, 2008.

[16] J. Dibiase, H. Silverman, and M. Brandstein, *Microphone Arrays. Robust Localization in Reverberant Rooms*, Springer, New York, NY, USA, 2001.

[17] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 288–292, 1997.

[18] C. Canton-Ferrer, J. R. Casas, M. Pardàs, and R. Sblendido, "Particle filtering and sparse sampling for multi-person 3D

tracking," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '08)*, pp. 2644–2647, October 2008.

[19] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[20] P. Salembier and L. Garrido, "Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval," *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 561–576, 2000.

[21] X. Giro and F. Marques, "Composite object detection in video sequences: application to controlled environments," in *Proceedings of the 8th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS '07)*, pp. 1–4, June 2007.

[22] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '99)*, pp. 246–252, June 1999.

[23] R. Rifkin and A. Klautau, "In defense of One-Vs-All Classification," *The Journal of Machine Learning Research*, vol. 5, pp. 101–141, 2004.

[24] M. Chan, Y. Zhang, and T. Huang, "Real-time lip tracking and bi-modal continuous speech recognition," in *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, 1998.

[25] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *Proceedings of the IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, 2005.