Journal version available at http://dx.doi.org/10.1016/j.patrec.2015.06.026

Pattern Recognition Letters
journal homepage: www.elsevier.com

Reliable Detection of Audio Events in Highly Noisy Environments

Pasquale Foggia[a], Nicolai Petkov[b], Alessia Saggese[a], Nicola Strisciuglio[a], Mario Vento, *IAPR Fellow*[a,**]

[a]*Dept. of Information and Electrical Engineering and Applied Mathematics, University of Salerno, via Giovanni Paolo II, 84084 Fisciano (SA), Italy*
[b]*Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, Nijenborgh 9, 9747 AG Groningen, The Netherlands*

ABSTRACT

In this paper we propose a novel method for the detection of audio events for surveillance applications. The method is based on the *bag of words* approach, adapted to deal with the specific issues of audio surveillance: the need to recognize both short and long sounds, the presence of a significant noise level and of superimposed background sounds of intensity comparable to the audio events to be detected. In order to test the proposed method in complex, realistic scenarios, we have built a large, publicly available dataset of audio events. The dataset has allowed us to evaluate the robustness of our method with respect to varying levels of the Signal-to-Noise Ratio; the experimentation has confirmed its applicability in real world conditions, and has shown a significant performance improvement with respect to another method from the literature.

1. Introduction

Audio analysis has been traditionally focused on the recognition of speech (Anusuya and Katti, 2010; Besacier et al., 2014), speaker identification (Cordella et al., 2003; Chetty and Wagner, 2005; Saquib et al., 2010; Roy et al., 2012) and scene categorization (Cai et al., 2008; Pancoast and Akbacak, 2012). Recently, research in the area of intelligent surveillance systems shifted its attention to the automatic detection of abnormal or dangerous events through the analysis of audio streams acquired by microphones. Indeed, there are kinds of event (gun shots, screams and glass breakings) that can be effectively detected by using audio sensors but are much less evident by looking at the video stream. Audio analytics systems can be easily and inexpensively employed together with existing surveillance infrastructures, today mainly based on video analytics algorithms that use object tracking techniques (Di Lascio et al., 2013). Indeed, many IP surveillance cameras are already equipped or predisposed to be connected to a microphone, making possible a joint analysis of audio and video streams (Cristani et al., 2007).

One of the problems in audio surveillance applications is that the sounds of interest are superimposed on significant background sounds, often with very different values of the *signal to noise ratio* (SNR). Thus, it might be difficult to separate the noise to be ignored from the sounds to be recognized. Moreover, the properties of those events might be evident at different time scales: for instance, a gun shot is an impulsive sound and its spectrum distribution over time is very different from that of a scream that is an exemplary sustained sound.

The state-of-the-art audio surveillance methods (see Crocco et al. (2014) for a comprehensive review) can be categorized in two main groups, depending on the architecture employed for classification. In the first group, the approach is to extract characteristic features (Mel-Frequency Cepstral Coefficients or Wavelet-based coefficients) from small audio frame in which the input signal is divided and use them in combination with a classifier to take decisions. Vacher et al. (2004) and Clavel et al. (2005) employ Gaussian Mixture Model (GMM) classifiers trained on different sets of features in order to detect screams or gun shots, while Valenzise et al. (2007) use them to address the problem of modeling the background sounds. In order to reduce the influence of the background sounds on the classification results, Rabaoui et al. (2008) adopted a pool of One Class Support Vector Machines (OC-SVM) with a novel dissimilarity measure. Performing only a short-time analysis, these methods display limited capabilities when confronted with both sustained and impulsive sounds, and a low robustness to background sound variations.

In the second group, more complex architectures have been proposed to increase the reliability of the systems. Rouas et al.

**Corresponding author: Tel.: +39 089 963006
e-mail: mvento@unisa.it (Mario Vento, *IAPR Fellow*)

(2006) propose an approach that combines GMM and Support Vector Machine (SVM) classifiers for detecting screams in outdoor environments, together with an adaptive thresholding on sound intensity for limiting the number of false detections. Ntalampiras et al. (2009) propose a two-stage GMM classifier in which the first stage aims to separate normal and abnormal sounds while in the second stage they are classified in one of the classes of interest. Conte et al. (2012) present a method in which impulsive and sustained sounds are analyzed by means of two classifiers that work at different time scales. The method uses a quantitative estimation of the reliability of each classification to reduce the false detections by rejecting the classifications that are not considered sufficiently reliable. The temporal sequence of symbols that represent spectral shapes has been taken into account by Chin and Burred (2012), that classify the audio events by matching sub-sequences of the reference events using the Genetic Motif Discovery technique. The event detection task is formulated by Foggia et al. (2014) as an object detection problem in the Gammatone image of the sound.

Generally, more complex classification architectures require a ground truth defined both at short- and long-time level, increasing the human labor time needed to label the data set. Complex architectures achieve stronger robustness to the background noise, but require higher computational resources. When, instead, a high-level representation of the data is used, the discriminative power of the systems improves while the classification scheme is kept simple.

In this paper we present a system for audio analysis based on the *bag of words* approach, as an extension of the paper by Carletti et al. (2013). The bag of words paradigm has been successfully applied in other fields, ranging from textual documents retrieval (Joachims, 1998), to human actions recognition (Foggia et al., 2013), video-based object detection (Sivic and Zisserman, 2009) or music classification (Fu et al., 2011).

The application of this paradigm to audio has been pioneered by Pancoast and Akbacak (2012). The underlying idea is that the audio stream can be thought of as being composed of small perceptual units of hearing, which we call *aural* words, whose distribution over a finite interval of time allows to characterize the type of sound. While a single aural word describes the short-time characteristics of the audio signal, the presence of certain words together is likely representative of the occurrence of a given event. In the paper of Pancoast and Akbacak (2012), this paradigm is applied to the classification and indexing of multimedia assets: namely, the authors classify a set video clips into different kinds of scenes on the basis of their audio content. That problem is significantly different from audio surveillance, for the following reasons: the scene to be recognized is long at least several seconds, while in surveillance the sound of interest can be very short (e.g. a gunshot can last for less than 200 milliseconds); the audio quality is usually good, while audio surveillance has to deal with noise introduced during both the acquisition (e.g. because of low quality microphones and of distance) and the transmission of the sound (e.g. compressed transmission over a low bandwidth network); the whole scene has to be recognized, while in a surveillance scenario, instead, the event of interest is one of many sounds simultaneously present in the environment, and not necessarily the loudest one, since other background sounds can be produced by sources that are closer to the microphone.

The specific characteristics of the surveillance problem strongly impact on how the bag of word approach must be tailored in order to be effectively applied. First, the fact that the sounds of interest will usually occur superimposed with other background sounds must be considered during the system training. Second, the high level of noise means that an exact matching of the aural words may fail, giving erroneous detections.

In this paper we present a system for audio surveillance based on a specific adaptation of the bag of words approach. The system has been validated on a large data set introduced in this paper, especially designed for benchmarking in realistic environmental conditions, and made publicly available (*http://mivia.unisa.it/*). The main contributions and the differences of this work with respect to Carletti et al. (2013) are: *a*) the design and realization of a wide and challenging data set of sounds, with highly noisy background, occurring at different SNRs ranging from $5dB$ to $30dB$; *b*) a technique for reducing the required training time without affecting the accuracy; *c*) an improvement of the robustness to noise through the use of soft assignment to bags, to limit the errors due to the exact matching of aural words; *d*) a detailed analysis of the robustness of the proposed approach to significant variations of the SNR and with very different background sounds. The newly proposed data set is composed by 6000 events (glass breakings, gun shots and screams) that occur in various environmental conditions.

2. The proposed method

Given M classes of events of interest $C_1, \ldots, C_M$, and a class of background sounds C_0 the system has to detect if and when a certain audio event occurs and effectively distinguish it from the background sounds. The audio stream is first divided in small frames of the order of milliseconds for which short-time, low-level features, that we call *aural words*, are computed and then used to construct a higher-level feature vector whose elements are indicators of the occurrence of such short-time features. The set of words is obtained by means of a clustering process that quantizes the original space of short-time features. The detection of audio events is performed in a time window of m seconds that moves along the audio stream. For each time window, the histogram of the occurrences of the aural words is constructed and is used as a feature vector to be fed to a pool of SVM classifiers, one for each class C_i. In the operating phase, the decisions taken independently by each SVM are combined together to obtain the output class of the sound.

2.1. Short-time and long-time descriptors

In contrast to video signals, in which a scene can persist even for several seconds, an audio signal might show huge temporal variations within a few milliseconds. Thus, in order to take into account the short-time variability, the input audio stream is first segmented into groups of N partially overlapping frames of duration T_F, windowed by a Hamming window. The choice of T_F is influenced by two contrasting effects: if the value is too short,

Category	Features
Spectral features	spectral centroid, spectral spread, spectral rolloff, spectral flux
Energy features	energy, 4 sub-bands energy ratios, volume
Temporal features	Zero-crossing rate (ZCR)

Table 1: Feature set used to build the short-time descriptor.

the frame will be unable to accurately represent low-frequency components of the sounds. Conversely, if it is too long the frame will not represent adequately short-time changes in the audio signal. We found as a reasonable compromise for a reliable analysis a value of $T_F = 32msec$ for audio streams sampled at $32KHz$. Every frame is built by advancing the frame window by $T_F/4$ and contains $L = 1024$ PCM samples. A set of spectral and temporal features (Peeters, 2004) and energy features (Liu et al., 1998), listed in Table 1, are used to build the short-time descriptor. A complete explanation with mathematical formulations is reported in Carletti et al. (2013). In the same way a text cannot be classified from a single word but rather from the occurrence of different words, our hypothesis is that a given audio event is characterized by the occurrence of specific basic sound units. In order to derive a finite set of atomic sounds, which we call *aural words* to emphasize the fact that they are related to perceptual units of hearing and not to linguistic units, we quantized the space of the short-time descriptors using the K-Means clustering algorithm. Since the space of short-time descriptors is dense, an uniform down-sampling of the space, by a factor 2, allowed to considerably speed-up (up to 90% of time less) the clustering process, which ended up in a set of K well-representative clusters, without influencing the final performance of the system. The centroids of the K obtained clusters are collected in the set of words w_i that constitutes the dictionary $W = \{w_1, \ldots, w_K\}$ of the system. Each vector w_i in the dictionary is representative of a recurrent atomic sound unit, whose occurrence increases the statistical evidence of being in presence of a given sound. Of course it is expected that a single word is not representative of the presence of an event of interest; thus we performed the detection at a time scale of the order of seconds by looking at the occurrences of certain aural words that are distinctive for the sounds of interest. Since the K-Means algorithm only requires unlabeled samples, for the training of the proposed system it is not strictly necessary to have a ground truth with a granularity of a single frame. It is, instead, sufficient to define the true labels, which indicate the presence or not of a specific event, only at a longer time scale, corresponding to time windows of the order of seconds. Thus, it is a significant advantage with respect to methods that require a ground truth at frame level, so reducing the human labor time.

In the conventional bag of words approach, the construction of the descriptors uses the so called *hard assignment* technique: for each feature vector v_i, the dictionary is searched for the closest word w_j to v_i. Finally the long-time descriptor $H = (h_1, \ldots, h_K)$ is calculated as the histogram of the occurrences of the different words, as in the following:

$$h_j = \sum_{i=1}^{N} a_{ij}, \quad j = 1, \ldots, K, \tag{1}$$

where a_{ij} is an indicator function with value 1 if the closest word to v_i is w_j:

$$a_{ij} = \begin{cases} 1 \text{ if } i = \arg\min_j D(v_i, w_j), & j = 1, \ldots, K \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

with $D(v_i, w_j)$ a distance measure between the i-th vector in the time window and the j-th word of the dictionary (for uniformity with the distance metric employed in the K-Means algorithm, the Euclidean distance was used).

While hard assignment may work well in contexts with low to moderate noise, the effect of noise on the vectors v_i is somewhat amplified by the quantization introduced by this rule. Since in audio surveillance we expect the noise to be significant, in order to contrast this effect we have adopted the so called *soft assignment* described by Liu et al. (2011); Equation 2 is replaced by:

$$a_{ij} = \frac{\exp(-\beta D(v_i, w_j))}{\sum_l \exp(-\beta D(v_i, w_l))} \tag{3}$$

with the parameter β used to control the "softness" of the assignment.

It is worth noting that such representation is invariant with respect to the position of the ¡ event of interest within the considered time interval. Indeed, since the temporal arrangement of the aural words does not affect the construction of the histogram, a target sound that occurs at the beginning of an interval and another one that occurs at the end can be modeled with the same histogram, so contributing to the stability and simplicity of the representation.

2.2. The classifier

The long-time descriptors are used to train a pool of SVM classifiers (Cortes and Vapnik, 1995) with a ground truth defined at interval time-scale. The motivation of this choice lies in the ability of a SVM classifier, like other classifiers based on discriminant analysis, to construct a decision function that gives only to a subset of the features a non-zero weight. More in details, it means that the classifier learns which aural words are really discriminative for a certain class of events ignoring the others. We have used the original, linear version of the SVM, and not the kernelized one, since it provided satisfactory results in our experiments. Since the proposed system has to face a multi class classification problem and the SVM is essentially a binary classifier, we adopted the 1-vs-all SVMs classification scheme (Fig. 1e). Namely, we have a pool of $M + 1$ 1-vs-all SVM classifiers (where M is the number of the classes to be recognized). The i-th classifier (with $i = 0, \ldots, M$) is trained using as positive examples the samples from class C_i and as negative examples all the samples from the other classes. We observed that, employing a SVM classifier also for the background sounds allows to reduce the detection of false positive events. During the operating phase, each input pattern (long-time descriptor) is fed to the pool of SVM classifiers; each classifier gives as its output a score s_i, which indicates its confidence, higher for more robust decisions. The final class C is assigned to the input pattern H

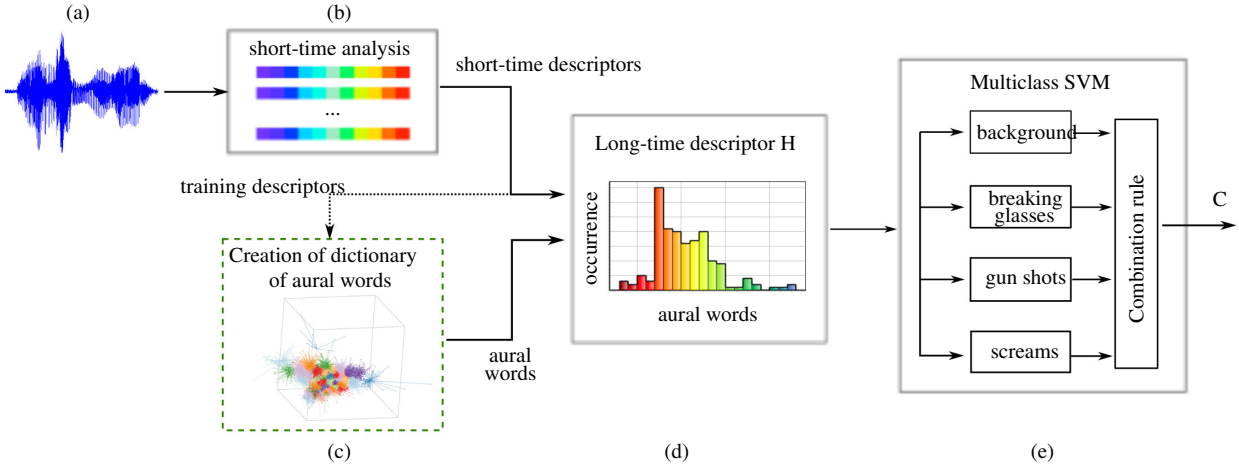

Fig. 1: Overview of the system architecture. The input audio signal (a) is divided in small frames (b) and a short-time descriptor is computed for each of them. A dictionary of aural words is created during the training phase (green box) by means of a clustering process (c). Then the histogram of the occurrences of the aural words in a m-seconds time windows (long-time descriptor H) is computed (d) and is fed to a multiclass SVM classifier (e) that performs the detection of events.

through the following combination rule:

$$C = \begin{cases} C_0, & if\ s_i < \tau\ \forall i = 0, \ldots, M \\ \arg\max_i s_i, & otherwise. \end{cases} \quad (4)$$

Namely, if at least one of the scores is higher than a threshold τ, the vector is assigned to the class whose SVM gives the maximum score. Otherwise, if all the classifiers give a score $s_i < \tau$, the vector will be assigned to the background class C_0. The overall architecture of the proposed system is depicted in Fig. 1. Only during the training phase, the short-time descriptors extracted from the input signals are used to build a dictionary of aural words (Fig. 1c), which serve then to compute the long-time descriptors (Fig. 1d). The SVM classifiers (Fig. 1e) learn which aural words are specific for a given class of sounds while through the combination of the scores of the SVMs, during the testing phase, the detection of events is performed.

3. Experimental results

The proposed system has been experimentally validated considering a typical application of audio surveillance in which three classes of audio events have to be detected: *scream*, *glass breaking* and *gun shot*.

A key requisite for an audio surveillance system is the ability to detect events of interest, even when they are mixed with different kinds of background sounds at different energy levels. Thus, the approach followed in other application domains of training a system by using a set of training samples each containing only one kind of sound (either an event of interest or a background sound) will give a poor performance during the test in realistic conditions.

In order to address this problem, and to provide a quantitative assessment of its impact, we have decided to construct a training set (and a test set) in which the individual sounds are not present isolated but are already superimposed to each other. After collecting a large number of audio clips, we combined them in several ways, obtaining an extended data set, so as to produce very challenging detection tasks, with low SNR and with the events mixed with a plurality of background noises.

To the best of our knowledge, there are no publicly available data sets for the benchmarking of audio surveillance applications. Thus, we constructed our own dataset of PCM audio clips sampled at 32 KHz and with a sample resolution of 16 bits. The data set, available at http://mivia.unisa.it, contains highly noisy environmental sounds with events of interest superimposed at different values of the SNR (in our case, 6 different values), making the detection and classification of events very challenging tasks. The intensity of the background sound is modulated in order to obtain low levels of SNR and simulate events that occur at various distances from the microphone. Originally, we collected a total of 650 audio clips, 271 of which relative to sounds that belong to the three classes of interest and the others relative to a wide variety of different sounds both from indoor and outdoor environments (silence and Gaussian noise, rain, whistles, crowded ambiance, vehicles, household appliances, bells, applauses and claps). Thus, even though the classes of interest are only three, the number of different types of sound the system must deal with is significantly higher; furthermore, some of the background sounds are similar to the classes of interest (e.g. the voices of people in a crowded ambiance are similar to screams). All the audio clips have been recorded with an Axis P8221 Audio Module and an Axis T83 omnidirectional microphone for audio surveillance.

The audio clips from the original data set have been normalized so that they have all the same overall energy and then they were split in two disjoint groups comprising the 70% and 30% of the total amount of sounds from the original set, respectively. We used the clips from the first group to build the training set and the ones from the second set to build the test group. The procedure described in the following was applied both for the training and the test set.

First, the sound $B(n)$ of a complex environment is created by mixing a randomly defined number $d \in \{1, 2, 3\}$, of the above mentioned background sounds, as follows:

$$B_j(n) = \sum_{i=1}^{d} b_i(n), \quad (5)$$

where $b_i(n)$ are the d background sounds used to create the com-

Table 2: Summary of the composition of the data set.

Data set description

	Training set		Test set	
	#Events	**Duration (s)**	**#Events**	**Duration (s)**
BN	-	58371.6	-	25036.8
GB	4200	6024.8	1800	2561.7
GS	4200	1883.6	1800	743.5
S	4200	5488.8	1800	2445.4

plex environmental sound. All the audio files in the created data set have a duration of about 3 minutes; in case an original background sound has a duration shorter than 3 minutes, it is replicated in order to fit the established length.

Once the environmental sound has been created, a number N_e of foreground events is randomly chosen from the original data set and superimposed to the environmental sound, in order to simulate the occurrence of an event in a real and complex environment. In this way, an event can be present in the final data set a plurality of times, but every time it appears with a different background noise. Moreover, since in real situations the source of a target event can be at different distances from the microphone, different values of the SNR of each event have been produced in the creation of the final data set. In particular, when a foreground sound is mixed with the environmental sound, the energy of the foreground sound is amplified or attenuated according to a specific value of the SNR, SNR^p with $p = \{5dB, 10dB, 15dB, 20dB, 25dB, 30dB\}$, for the target sound. The rule for the construction of the audio event $y_j^p(n)$ at a certain SNR value is defined as follow:

$$y_j^p(n) = \sum_{i=1}^{N_e} \{B_j(n) \oplus_{[s_i,e_i]} A_p \overline{x}_i(n)\}, \tag{6}$$

where

$$A_p = 10^{SNR^p/20} \frac{rms(B_j(n))}{rms(\overline{x}_i(n))}. \tag{7}$$

The amplification (or attenuation) coefficient A_p depends on the specific SNR^p value and on the root mean square values (rms) of the environmental sound and of the foreground sound. With $\oplus_{[s_i,e_i]}$ we define an operator that mixes the signal $A_p\overline{x}_i(n)$ with the signal $B_j(n)$ in the interval delimited by $[s_i, e_i]$, starting and ending points of the target sounds respectively.

The final data set consists of a training set and a test set that contain, respectively, 396 and 184 audio files of about 3 minutes, each of them containing a sequence of events at a specific SNR value. The total duration is about 20 hours for the training set and about 9 hours for the test set, making the database huge. A total of 6000 events per class have been collected, 4200 in the training set and 1800 in the test set. In the following we will refer to the different classes of events with the abbreviations *GB* for glass breaking, *GS* for gunshot, *S* for scream and BN for background noise. In Table 2 a summary of the composition of the data set is reported.

3.1. Performance evaluation

For evaluating the algorithm performance, we considered two measures: the recognition rate of the events of interest and the false positive rate (FPR), i.e events of interest detected when only background sound is present.

Table 3: Results of the proposed system on the test set.

Training with isolated sounds

		Guessed class			
		GB	**GS**	**S**	**Miss**
True class	**GB**	84.3%	0%	0.1%	15.6%
	GS	29.6%	0%	1.7%	68.7%
	S	22.4%	0%	28%	49.6%

Proposed method – Hard assignment

		Guessed class			
		GB	**GS**	**S**	**Miss**
True class	**GB**	93.6%	0.2%	0.2%	6%
	GS	3.3%	81.6%	0.5%	14.6%
	S	2.8%	0.9%	79.3%	17%

Proposed method – Soft assignment

		Guessed class			
		GB	**GS**	**S**	**Miss**
True class	**GB**	94.4%	0.2%	0.2%	5.2%
	GS	3.5%	84.9%	0.5%	11.1%
	S	2.6%	0.9%	80.8%	15.7%

An event is considered as correctly detected if it is detected in at least one of the sliding time windows that overlap with it occurrence. Table 3 shows the classification matrix for three versions of the system: the first one is trained using isolated sounds; the second one uses the mixed training sounds with hard assignment, and the last one uses the mixed sounds with soft assignment. As it can be seen, the use of mixed sounds in the training is essential for achieving good results on the test set. Soft assignment also yields a significant improvement, especially for the gun shot class, which is more sensitive to errors in the codeword assignment because its samples are very short. In the rest of the discussion, we will refer to the soft assignment version, unless otherwise specified.

The average correct classification rate of foreground events, at different SNR values, achieved on the whole test set is 86.7%. The classification matrix also shows that most of the errors are mainly directed to the background noise class (missed detections), while confusion between different classes of interest is low. We count a false positive (FP) when an event of interest is erroneously detected in a time window that contains only background noise; it is worth pointing out that if in two consecutive time windows a foreground event is detected, we count, as it should be, a single false positive occurrence. Thus, the FPR is computed as the ratio of the detected false positive events to the total number of intervals between two foreground sounds. For the whole test set we achieved a FPR equal to 2.1%, being 0.83% false detected glass breakings, 0.74% gun shots and 0.53% screams. It is worth noting that more than 70% of the number of FP is concentrated in about one hour of background sound, composed mainly of household appliance and rain sounds, while the remaining 30% is distributed along more than 6 hours of other background sounds. In real conditions, it is difficult to have, in the same environment, both rain and

Table 4: Detailed results achieved by using the proposed classifier for different values of the SNR of the foreground sounds.

Proposed method - Detailed results				
SNR	**Recognition**	**Miss**	**Error**	**False Positive**
5dB	81.1%	12%	6.9%	11.5%
10dB	85%	12.1%	2.9%	2.4%
15dB	87%	10.9%	2.1%	1.3%
20dB	88.4%	9.9%	1.7%	1.2%
25dB	88.7%	9.9%	1.4%	1.2%
30dB	90%	9.2%	0.8%	1%
Average	86.7%	10.7%	2.3%	2.6%

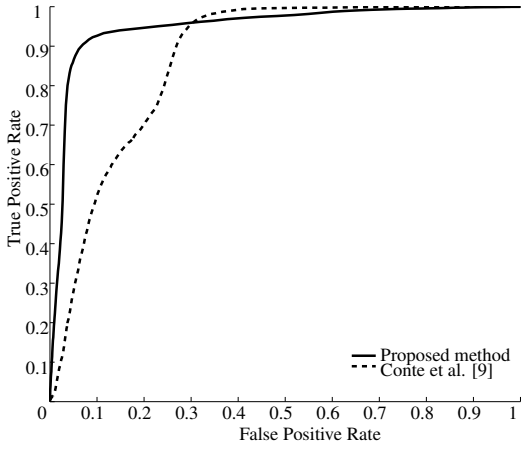

Fig. 2: ROC curve of the proposed system (solid line) compared with that of Conte et al. 2012 (dashed line).

household appliance sounds and it is thus expected to achieve a lower number of false alarms. Moreover, the system could be optimized for particular kind of environments, for instance by using a pre-filter, in order to decrease the FPR.

A more detailed analysis of the performance of the system on audio clips with different SNR values of the foreground events is reported in Table 4. As clearly expected, when the sounds of interest have higher values of SNR (and thus a lower level of the noise), the influence of the background noise is reduced leading to an improvement of the recognition rate and a reduction of FPR and miss rate. It is also worth noting that the recognition rate for events at 5dB SNR is only 5% lower than the average recognition rate on the whole data set and about 9% lower than the best value achieved for the events at 30dB SNR. The correct classification rate and the false positive rate achieved in different SNR conditions prove that the proposed system is robust to high background noise variations.

We have performed the analysis of the performance versus SNR also for the bag of words implemented with hard assigment. The result is that the increment of performance due to soft assignment is roughly equivalent to a 5dB increment of the SNR (e.g. soft assigment at 10dB of SNR has approximately the same performance of hard assignment at 15dB).

3.2. Performance comparison

We compared the performance of the proposed system with the ones achieved by the method of Conte et al. (2012), which employs a single-level representation and aggregates short-time decisions, taken by a LVQ classifier trained with the same short-time descriptors, at time window level using a rejection rule. The system attributes an interval to the class C_i that obtains the highest score $z_i = (n_i - \hat{n}_i)/\hat{n}_i$, where n_i is the number of frames in the interval assigned to the class C_i; $\hat{n}_i$ is a threshold that indicates a limit under which the i-th class is not considered as a candidate for the final decision. An interval is rejected (classified as background) if $z_i < 0$ for $\forall i = 1, ..., M$.

Table 5: Comparison of the proposed method with Conte et al. (2012) in terms of AUC for the foreground classes.

	Proposed Method	**Conte et al. (2012)**
GB	**0.954**	0.872
GS	**0.968**	0.886
S	**0.966**	0.938

Table 6: Results achieved by Conte et al. (2012) on the data set.

Conte et al. (2012) - Classification matrix					
		Guessed class			
		BG	**GS**	**S**	**Miss**
True class	**BG**	91.3%	5.3%	1.4%	1.9%
	GS	12.1%	80.6%	3.9%	3.4%
	S	7.6%	7.9%	79.8%	4.7%

First, we compare the performance of the two methods by using the receiver operating characteristic (ROC) curves that give an overall evaluation of the classification performance. The ROC curves, obtained varying the values τ for the proposed method and n_i for Conte et al. (2012) are depicted in Fig. 2. The proposed method clearly outperforms the other one, as the corresponding curve is closer to the left and top borders of the quadrant. We consider the area under the ROC curves (AUC), which is equal to 1 for a perfect classification, as a measure of the performance of the two methods and report the results in Table 5. The higher this measure, the better the overall performance of the system is. We observe that the proposed method (solid line) generally outperforms Conte et al. (2012) (dashed line), achieving an average AUC that is about 7.2% higher.

The introduction of a second level of representation exploits the long-time properties of the signal and contextual information about the environmental sounds, limiting the effects of the background noise on the detection of events. Due to the combination of short- and long-time analysis, the performance is generally better than a method based on a single representation level that considers for the evaluation only the short-time properties of the audio signal. In fact, when a decision is taken only at a lower level, like in the case of Conte et al. (2012), the reliability of the system is penalized by the effect of the background noise. Thus, the system can be useful for audio surveillance due to its robustness to the environmental noise and the consequently lower false alarm rate, even at low SNR.

Finally, in order to compare the performance of the two systems in operating conditions, we determined the value of the threshold $\tau = 0$ for the *bag of aural words* classifier and the values of $\hat{n}_{BN} = 266$, $\hat{n}_{GB} = 95$, $\hat{n}_{GS} = 26$, $\hat{n}_S = 62$ for Conte et al. (2012), through a cross-validation on the training set. In Table 6, we report the recognition results achieved by Conte et al.

(2012) on the proposed data set. The average recognition rate is 83.9%, that is lower than the one (86.7%) achieved by the proposed method. However, the difference is much more consistent for low values of the SNR: at $5dB$, the recognition rate of Conte et al. (2012) is 71.4%, about 10% less than the one of the proposed method, and the False Positive Rate is about 10% higher. This is evident from the graphs reported in the supplementary materials of the paper, in which the performance achieved by the two systems are reported for different values of SNR. These results show a higher robustness of the proposed system to the background noise with respect to Conte et al. (2012).

We conclude this section with some information on the computational cost of the proposed method. The processing time of the training phase is significant: on a 2.6GHz Opteron processor, the preparation of the codebook requires about 7 hours (reduced from about 70 hours needed before we introduced the technique described in Section 2.1); the training of the pool of SVMs requires about 2.5 hours. Once the system is trained, the execution of the algorithm is quite fast: on the processor used for the training, the execution in real time at a 32 KHz sound sampling rate requires about 3% of the time of a single CPU core. The system also runs in real time on an embedded Raspberry Pi board, making its deployment very inexpensive.

4. Conclusions

In this paper we proposed a system based on the *bag of aural words* approach for the detection of events in audio streams for surveillance applications. The bag of word approach has been tailored and adapted to the specificity of the application domain, such as the high noise level and the need to deal with loud background sounds superimposed to the events of interest. We experimentally validated the system on a large and challenging audio data set that we have made publicly available for benchmarking purposes. The performance results, compared with a state of the art approach (Conte et al., 2012), confirm the robustness of the proposed method with respect to the background noise and its applicability to real environments. In general, the main advantage of the proposed system is that the bag of words approach intrinsically takes into account contextual information about the environment to build a model of the events of interest and learns which features of the sound are distinctive for a specific class of events. This leads to more reliable performance and to a higher robustness to the background noise, even for highly noisy environments and low SNR levels.

5. Acknowledgements

This research has been partially supported by A.I.Tech s.r.l. (http://www.aitech-solutions.eu).

References

Anusuya, M.A., Katti, S.K., 2010. Speech recognition by machine, a review. CoRR abs/1001.2267.

Besacier, L., Barnard, E., Karpov, A., Schultz, T., 2014. Automatic speech recognition for under-resourced languages: A survey. Speech Commun. 56, 85–100.

Cai, R., Lu, L., Hanjalic, A., 2008. Co-clustering for auditory scene categorization. IEEE Trans. Multimedia 10, 596–606.

Carletti, V., Foggia, P., Percannella, G., Saggese, A., Strisciuglio, N., Vento, M., 2013. Audio surveillance using a bag of aural words classifier, in: IEEE AVSS, pp. 81–86.

Chetty, G., Wagner, M., 2005. Investigating feature-level fusion for checking liveness in face-voice authentication, in: Proc. of the 8th International Symposium on Signal Processing and its Applications, ISSPA-2005, pp. 66–69.

Chin, M., Burred, J., 2012. Audio event detection based on layered symbolic sequence representations, in: IEEE ICASSP, pp. 1953–1956.

Clavel, C., Ehrette, T., Richard, G., 2005. Events detection for an audio-based surveillance system, in: ICME, pp. 1306 –1309.

Conte, D., Foggia, P., Percannella, G., Saggese, A., Vento, M., 2012. An ensemble of rejecting classifiers for anomaly detection of audio events, in: IEEE AVSS, pp. 76–81.

Cordella, L., Foggia, P., Sansone, C., Vento, M., 2003. A real-time text-independent speaker identification system, in: ICIAP, pp. 632–637.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297.

Cristani, M., Bicego, M., Murino, V., 2007. Audio-visual event recognition in surveillance video sequences. IEEE Trans. Multimedia 9, 257–267.

Crocco, M., Cristani, M., Trucco, A., Murino, V., 2014. Audio surveillance: a systematic review. CoRR abs/1409.7787. URL: `http://arxiv.org/abs/1409.7787`.

Di Lascio, R., Foggia, P., Percannella, G., Saggese, A., Vento, M., 2013. A real time algorithm for people tracking using contextual reasoning. CVIU 117, 892–908.

Foggia, P., Percannella, G., Saggese, A., Vento, M., 2013. Recognizing human actions by a bag of visual words, in: Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conf. on, pp. 2910–2915.

Foggia, P., Saggese, A., Strisciuglio, N., Vento, M., 2014. Cascade classifiers trained on gammatonegrams for reliably detecting audio events, in: IEEE AVSS, pp. 50–55. doi:`10.1109/AVSS.2014.6918643`.

Fu, Z., Lu, G., Ting, K.M., Zhang, D., 2011. Music classification via the bag-of-features approach. Pattern Recognition Letters 32, 1768 – 1777.

Joachims, T., 1998. Text categorization with suport vector machines: Learning with many relevant features, in: Proc. of the 10th European Conf. on Machine Learning, Springer-Verlag. pp. 137–142.

Liu, L., Wang, L., Liu, X., 2011. In defense of soft-assignment coding, in: Computer Vision (ICCV), 2011 IEEE Int. Conf. on, pp. 2486–2493.

Liu, Z., Wang, Y., Chen, T., 1998. Audio Feature Extraction and Analysis for Scene Segmentation and Classification. The Journal of VLSI Signal Processing 20, 61–79.

Ntalampiras, S., Potamitis, I., Fakotakis, N., 2009. An adaptive framework for acoustic monitoring of potential hazards. EURASIP J. Audio Speech Music Process. 2009, 13:1–13:15.

Pancoast, S., Akbacak, M., 2012. Bag-of-audio-words approach for multimedia event classification, in: Proc. of the Interspeech 2012 Conf.

Peeters, G., 2004. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Tech. Rep.. IRCAM.

Rabaoui, A., Davy, M., Rossignol, S., Ellouze, N., 2008. Using one-class svms and wavelets for audio surveillance. IEEE Trans. Inf. Forensics Security 3, 763–775.

Rouas, J.L., Louradour, J., Ambellouis, S., 2006. Audio events detection in public transport vehicle, in: IEEE ITSC, pp. 733–738.

Roy, A., Magimai-Doss, M., Marcel, S., 2012. A fast parts-based approach to speaker verification using boosted slice classifiers. IEEE Trans. Inf. Forensics Security 7, 241–254.

Saquib, Z., Salam, N., Nair, R., Pandey, N., Joshi, A., 2010. A survey on automatic speaker recognition systems, in: Signal Processing and Multimedia. Springer Berlin Heidelberg. volume 123 of *Communications in Computer and Information Science*, pp. 134–145.

Sivic, J., Zisserman, A., 2009. Efficient visual search of videos cast as text retrieval. IEEE Trans. Pattern Anal. Mach. Intell 31, 591–606.

Vacher, M., Istrate, D., Besacier, L., Serignat, J.F., Castelli, E., 2004. Sound Detection and Classification for Medical Telesurvey, in: ACTA Press, C. (Ed.), Proc. 2nd ICBME, Innsbruck, Austria. pp. 395–398.

Valenzise, G., Gerosa, L., Tagliasacchi, M., Antonacci, F., Sarti, A., 2007. Scream and gunshot detection and localization for audio-surveillance systems, in: IEEE AVSS, pp. 21–26.