This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI 10.1109/TITS.2015.2470216, IEEE Transactions on Intelligent Transportation Systems

Audio surveillance of roads: a system for detecting anomalous sounds

Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio and Mario Vento *IAPR Fellow*

***Abstract*—In the last decades several systems based on video analysis have been proposed for automatically detecting accidents on the roads so as to ensure a quick intervention of emergency teams. However, in some situations the visual information is not sufficient or sufficiently reliable, while the use of microphones and audio event detectors can significantly improve the overall reliability of surveillance systems. In this paper we propose a novel method for detecting road accidents by analyzing audio streams so as to identify hazardous situations like tire skidding and car crashes. Our method is based on a two layer representation of the audio stream: at a low level, the system extracts a set of features able to capture the discriminant properties of the events of interest; a high level representation based on the bag of words approach is then exploited in order to detect both short and sustained events. The deployment architecture for using the system in real environments is discussed, together with an experimental analysis carried out on a data set made publicly available for benchmarking purposes. The obtained results confirm the effectiveness of the proposed approach.**

***Index Terms*—Hazard detection, accident detection, audio events, audio detection, tire skidding, car crashes.**

I. Introduction

In the last years, a need for more security and safety in public environments has risen due to the increasing number of people and transportation vehicles that move around cities. Road traffic monitoring involves, for instance, the detection of accidents or road disruptions to quickly ensure the intervention of emergency teams and to guarantee the safety of the people [1]. In fact, it has been shown [2], [3] that the reduction of the time between the moment in which an accident occurs and the moment in which the emergency team is dispatched substantially decreases the mortality rate (approximately by 6%). Within this context, cameras have been widely used to control the behavior of vehicles by tracking their trajectories [4]–[7] near traffic lights or in proximity of road crosses in order to detect abrupt maneuvers, or on motorways to monitor the traffic flow and detect long queues [8], [9].

However, in certain cases, the visual information is not sufficient to reliably understand the activity of vehicles or to detect possibly hazardous situations. For instance, a tire skidding on the road has a very distinctive acoustic signature that is not detectable from video streams but can be an evidence of an anomalous situation (an accident or a dangerous state of the road) that requires human intervention to ensure safety. Furthermore, the abnormal events can happen outside the field of view of the camera, making it impossible to be detected both by a human operator and by an automatic video analytics system. In such cases, the use of microphones and the processing of the audio stream as a complementary tool to the video analysis may improve the detection abilities of security systems [10], [11] and, in general, the reaction time of the emergency teams. As a matter of fact, nowadays, IP cameras used for surveillance are normally equipped with embedded microphones that facilitate the deployment of audio analysis systems.

One of the main advantages of audio analysis systems is that they do not have to deal with variations in illumination conditions and can be equally employed during day and night. However, the problem of detection of audio events in open environments is very challenging: one of the main issues is that the events of interest are superimposed to a significant level with background noise; furthermore, it is difficult to model a priori all the possible background sounds that may occur in road environments. Think, for example, about a very busy highway where an accident occurs: an audio event detector needs to be able to separate the background noise due to the vehicle flow from the car crash (the event of interest) potentially occurring at a significant distance from the microphone. In such a case, the signal to noise ratio (SNR) is very low, thus making the recognition of such events a very complex task. Another typical problem the audio analysis systems has to face is related to the duration of the events of interest: a tire skidding, for instance, is typically a sustained sound and may last several seconds, while a car crash is an impulsive sound and its duration is very limited in time.

In the last decades a large number of methods dealing with the analysis of audio streams has been proposed, ranging from speech recognition [12], [13] and scene classification [14], [15] to speaker identification [16], [17]. More recently, a growing interest for audio analysis has been also shown in surveillance applications, in order to detect crimes for public transport security [18]–[20], the maximum speed of vehicles for security reasons [11], [21], [22] or accidents on the roads [2], [3], [23].

In this paper, we focus on the problem of road surveillance, and we propose a system tailored for the automatic detection of two hazardous situations, namely tire skidding and car crashes, by analyzing the sound captured by microphones. A comprehensive review of the state of the art approaches focusing on surveillance systems has been recently proposed

Pasquale Foggia, Alessia Saggese and Mario Vento are with DIEM, University of Salerno - Italy.

Nicola Strisciuglio is with the DIEM, University of Salerno - Italy and with Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen - The Netherlands.

Nicolai Petkov is with Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen - The Netherlands.

Manuscript received March XX, 2015; revised XXX XX, 2015.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI 10.1109/TITS.2015.2470216, IEEE Transactions on Intelligent Transportation Systems

Set	Type	Description	Ref.
1	Temporal and Spectral	• volume, energy, zero crossing rate • Spectral centroid, spectral spread, roll-off frequency, spectral flux • energy ratio in 4 sub-bands	[24]–[26]
2	Cepstral	• 13 Mel-frequency Cepstral Coefficients (MFCC)	[27]
3	Psychoacoustical	• Energy ratio in the first 24 critical bands of hearing	[28]

TABLE I: Details of the three low-level feature sets used for the experiments.

in [29], where it is highlighted that the detection of audio events can be considered as a traditional pattern recognition problem. In fact, the common idea is that the data to be analyzed is described by means of a set of features, whose values are used to form a vector representation of the pattern of interest. A feature is a salient characteristic of the pattern to be detected or classified, while a feature set aims at effectively describing patterns from different classes: similar patterns in the real world should have very close vectors in the feature space. The feature vectors are thus used to train a classifier, which creates models of the patterns of different classes through a learning process. Then it employs such models to classify newly observed patterns in the testing phase [30]–[37]. In the last years traditional classification schemes have been improved, and more sophisticated architectures have been proposed in order to increase the overall reliability of the audio detector [20], [38] or to take into account the different time resolution of the events of interest [39], [40].

Starting from a preliminary work [24], in this paper we present a detection system based on a high-level representation of the audio stream, able to take into account both the short- and long-time properties of the events of interest. Thanks to the use of a *bag-of-words* approach, our method learns which are the short-time characteristics of an event that are discriminant for such event on a longer time scale and that differentiate it from the background sound. This is a very important property, especially in the considered domain. In fact, in the case of the application at hand, a car crash sound is characterized by an abrupt variation of energy in time while a skidding tire is a sustained sound whose energy is concentrated in a narrow interval of frequencies.

We validated the system on a data set[1] that we made available for benchmarking purposes. In the proposed data set, the sounds of interest are not isolated but superimposed on different typical background sounds of roads and traffic jam, in order to consider the occurrence of such abnormal events in real-world conditions.

The paper is organized as follows: in Section II we describe the the proposed method and its rationale; in Section III we provide an overview of the system set up and an analysis on the positioning of the microphones; then, we present a detailed analysis of the performance in Section IV. Finally, we draw conclusions in Section V.

II. METHOD

The purpose of the proposed system is to distinguish audio events of interest from the background sound and classify them into one of M classes. The rationale of the proposed approach is based on the consideration that a sound is composed of small, atomic audio units, similarly to a text that is composed of a number of words, and the occurrence of particular units in a given time interval is an indicator of the presence of a certain event.

In order to build a description of the audio stream based on such assumption, a classification architecture exploiting the *Bag of Words* approach is employed. The bag of words technique has been widely applied for text categorization, in which the datum to be classified is represented by counting the occurrences of low-level features (*words*) and constructing a (high-level) vector whose dimensionality is equal to the number of possible words contained in a dictionary. The high-level vector corresponds, thus, to the histogram of occurrences of words, used for the classification of the text.

In the proposed architecture for audio analysis, the following layers have been defined: 1) extraction of low-level features, 2) learning of a dictionary of basic audio words, 3) construction of a high-level vector and 4) classification. Below, a detailed explanation of each layer is provided.

A. Low-level features extraction

In contrast with video streams, an audio signal can show abrupt variations within few milliseconds. Thus, in order to take into account its short-time variability, the audio stream is framed in small, partially overlapped, chunks (frames) of duration T_f. The value T_f has to be chosen to take into account the analysis of both low and high frequency components at the same time: with a very short frame, for instance, the system will not be able to consider low-frequency components; conversely, with a very long frame, high-frequency components will be averaged over a long time interval. For each frame, the system computes a vector of low-level features.

Three sets of low-level features have been considered and experimented with, namely the Mel-frequency cepstral coefficients (MFCC) [27], energy ratios in Bark sub-bands [28] and features based on temporal and spectral characteristics of the signal [25], [26], previously employed in [24]. More details on the three feature sets are reported in Table I.

B. Dictionary learning

The low-level feature space is continuous and theoretically infinite, thus not suitable for the detection of the presence of specific relevant atomic units of sounds (hereinafter *audio words*). In order to derive a finite set of audio words, we use the K-means algorithm, which clusters the vectors on the basis of their similarity. The output of the K-means algorithm is a

[1]The data set is available at the url http://mivia.unisa.it/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI 10.1109/TITS.2015.2470216, IEEE Transactions on Intelligent Transportation Systems

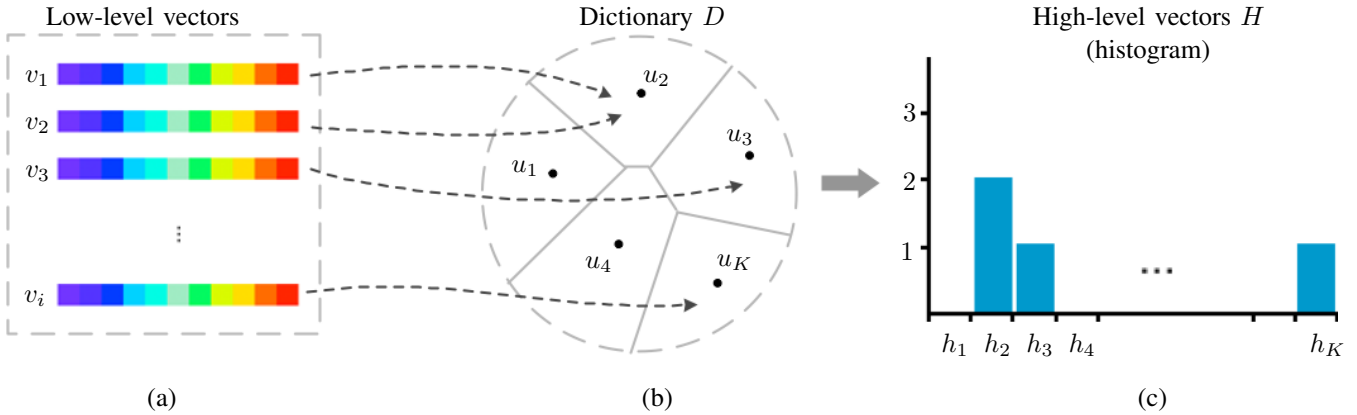


Fig. 1: Construction of the high-level representation. For each vector v_i the nearest audio word u_j in the dictionary is determined (b). Then, the occurrence counts of the single audio words are stored in a histogram, whose bins are h_j $(j = 1, \ldots, K)$ that constitutes the high-level vector (c). In the example, vectors v_1 and v_2 have u_2 as the nearest audio word in the dictionary. Thus, the second bin of the histogram has a value equal to 2. In the same way, audio word u_3 has only one close vector, resulting in a value equal to 1 for the third bin of the histogram. Audio words u_1 and u_4, instead, have no occurrences.

set of K points that correspond to the centroids of the clusters. Since a centroid is representative for a group of similar low-level vectors, we consider the set $D = \{u_1, \ldots, u_K\}$ of the centroids as the dictionary of basic audio words.

C. High-level representation

In Fig. 1, a sketch of the process of construction of the high-level representation is shown. Given the dictionary D, for each low-level vector v_i, the closest audio word u_j is determined. The occurrences of each word u_j in a time-limited interval are used to build a high-level feature vector. Such vector corresponds to the histogram $H = (h_1, \ldots, h_K)$, whose bins are computed as:

$$h_j = \sum_{i=1}^{N} \delta\left(b_i, j\right), \quad j = 1, \ldots, K, \tag{1}$$

where $\delta(\cdot)$ is the Kronecker delta and b_i is the index of a word within the set D, determined as:

$$b_i = \arg\min_j d(v_i, u_j), \quad j = 1, \ldots, K, \tag{2}$$

where $d(v_i, u_j)$ is a dissimilarity measure between the vector u_j and the prototype v_i (the Euclidean distance is considered).

D. Classification architecture

Our hypothesis is that certain classes of sounds are considered to have distinctive audio words that allow the system to differentiate such sounds from the other classes. A pool of $M + 1$ Support Vector Machines (SVM), each of them dedicated to the detection of a certain class of sounds (M events of interest plus the background sounds), has been trained with the high-level feature vectors. The SVM classifier is particularly suited for the employed sound representation since it is able to learn which are the words that are relevant for a particular class of events and discard those words that do not contribute to an effective classification, giving them a very low weight. We employed SVM with linear kernel, which gives satisfactory results in our experiments coupled with fast processing that is important for real-time responses.

The SVM is, originally, a binary classifier. Thus, a pool of SVM (Fig. 2) is realized in order to face the multi-class problem at hand. The i-th classifier is trained using as positive examples the samples from the class C_i and as negative examples all the samples from the other classes. During the testing phase, each classifier computes a score s_i which is a measure of the confidence of the classification, higher for more reliable decisions. The final class C is chosen as the one of the SVM that gives the highest score above a certain threshold λ:

$$C = \begin{cases} C_0, & if\ s_i < \lambda \quad \forall i = 0, \ldots, M \\ \arg\max_i s_i, & otherwise. \end{cases} \tag{3}$$

If all the classifiers give a confidence score $s_i < \lambda$ the time interval is classified as a background sound in class C_0. For our experiments the threshold is set as $\lambda = 0$. The use of a SVM classifier for the background class increases the robustness of the proposed system with respect to background noise and entails a reduction of false alarms.

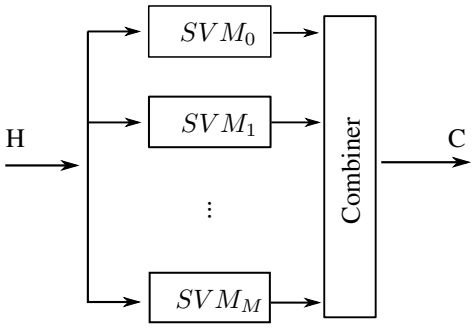


Fig. 2: Architecture of the classifier. The scores of the SVM classifiers are combined in order to determine the final class to be assigned to the input vector H.

III. Deployment Architecture

Our hypothesis for the deployment of the system is that we have a set $R = \{r_i | i = 1, ..., N_m\}$ of N_m microphones installed on one side of the road and located at a distance of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI 10.1109/TITS.2015.2470216, IEEE Transactions on Intelligent Transportation Systems

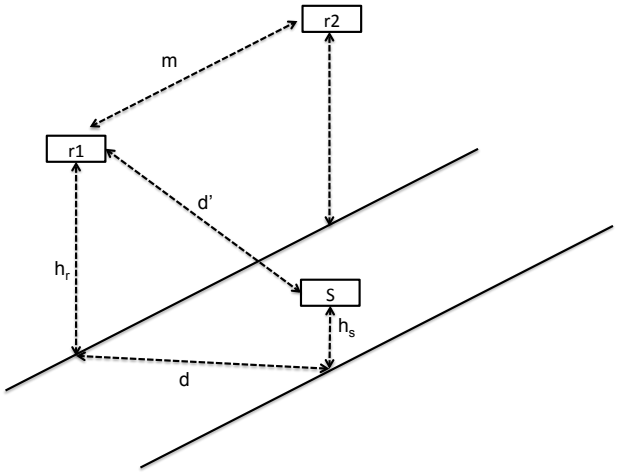


Fig. 3: A sketch of the deployment of the proposed system: a set R of microphones is located at a distance of m meters far from each other and at a height h_r. An event of interest can be recognized at a maximum distance of d meters from the closest microphone.

m meters far from each other and at a height of h_r meters (see Figure 3).

The choice of the distance m strongly depends on two factors: 1) the sound intensity of the events to be detected, 2) the maximum distance d from the microphone at which an event can still be detected by the system. Of course, d depends on the kind of environment the system has to work on: we expect that this value is higher for a country road (where only few vehicles go through the street with a low speed) than for a highway, where the number of vehicles and their speed are significantly higher.

In order to better understand the impact of the environment on the coverage capabilities of the microphones, we consider that the signal to noise ratio (SNR) of the sound acquired by a microphone (expressed in decibel) is computed as follows:

$$SNR = L_s(d) - L_n, \tag{4}$$

where $L_s(d)$ represents the intensity level expressed in decibel of the event of interest occurring at a distance d from the microphone, while L_n is the noise in decibel introduced by the traffic. In the following more information about the computation of these two contributions is provided.

A. Intensity level of the event of interest

Since the propagation of the sound is affected by spreading, absorption, ground configuration and so on, the intensity of the audio event acquired by the microphone is attenuated by a factor $A(d)$:

$$L_s(d) = L_s(d_0) - A(d), \tag{5}$$

where $L_s(d_0)$ is the sound intensity at a reference distance d_0.

According to the standard ISO 9613-2 [41], the attenuation can be computed as a combination of four contributions, which strongly depend on the environment where the sound is propagating:

$$A(d) = A_{div}(d) + A_{atm}(d) + A_{gr}(d) + A_{bar}(d). \tag{6}$$

Each of these contributions is determined by particular characteristics of the environment. In particular:

- A_{div} is due to the geometrical divergence; we suppose a spherical spreading from the source, whose sound is radiated equally in all directions; thus, the sound level is reduced by 6 dB for each doubling of distance from the source:
$$A_{div}(d) = 20 \log \frac{d}{d_0} + 11, \tag{7}$$
where 11, computed as $10 \cdot \log(4 \cdot \pi)$, is a constant that models the spherical spreading factor.
- A_{atm} is due to the atmospheric absorption during the propagation of the sound waves and can be computed as follows:
$$A_{atm}(d) = \frac{\alpha \cdot d}{1000}, \tag{8}$$
where α is the atmospheric attenuation coefficient, which is a function of the temperature, the humidity and the nominal frequency. According to [41], $\alpha = 32.8$ dB/Km assuming a temperature around 10° C and a nominal frequency of 4 kHz.
- The ground attenuation A_{gr} is the result of sound reflected by the ground surface interfering with the sound propagating directly from the source (the vehicle causing the sound of interest) to the receiver (the microphone). Let h_r and h_s be the receiver height and the source height, respectively. In order to compute A_{gr}, the standard [41] suggests to partition the area between the source and the receiver into three regions: *source region* (whose size is $30 \cdot h_s$), around the source, which determines the attenuation A_s; *middle region*, which determines the attenuation A_m; *receiver region* (whose size is $30 \cdot h_r$), around the receiver, which determines the attenuation A_r. A_{gr} is thus computed as:
$$A_{gr}(d) = A_s + A_m(d) + A_r. \tag{9}$$
In particular, at the nominal band of 4 kHz, A_r and A_s can be computed as follows:
$$A_r = A_s = 1.5 \cdot (1 - G) = 1.5. \tag{10}$$
According to the standard, the G value is equal to 0, since we suppose that the road is a *hard ground*. Conversely, A_m can be computed as:
$$A_m(d) = 3 \cdot q(d) \cdot (1 - G), \tag{11}$$
where
$$q(d) = \begin{cases} 0 & d \leq 30(h_s + h_r) \\ 1 - \frac{30(h_s+h_r)}{d} & d > 30(h_s + h_r) \end{cases}$$
- Finally, A_{bar} is due to the presence of barriers. Considering that the microphones are mounted directly on the road, this factor can be neglected in our scenarios.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI 10.1109/TITS.2015.2470216, IEEE Transactions on Intelligent Transportation Systems

B. Intensity level of the traffic noise

In the last decades the scientific community has proposed several approaches for modeling traffic noise, since it is considered very important in order to evaluate the acoustical impact both for environment management and urban planning. As shown in [42] and [43], there is not a commonly adopted rule but rather each country adopts its own standard: for instance, the CoRTN [44] procedure has been adopted in England, the RLS 90 [45] model in Germany, the C.N.R. model [46] in Italy and the NMPB in France [47].

A common idea of such methodologies is to take into account the traffic flow, both of light and heavy vehicles, the typology of the road surface and the distance between the microphone and the carriage generating the noise. In particular, in this paper we apply the CoRTN model in order to evaluate the traffic noise generated in different scenarios by taking advantage on the on line application provided by [48]. The CoRTN model evaluates the so called L_{10} (from now on L_n), that is the noise level exceeded for just 10% of the time over a period of one hour.

The main idea is to partition the road into a set of S segments (so as within one segment the noise level variation is lower than 2 dB) and to separately evaluate for each i-th segment the basic noise level L_i, taking into account attenuation due to the distance as well as the particular environment. Finally, the contribution of all the segments is combined so as to obtain the overall noise L_n.

According to the CoRTN model, the noise L_i for the i-th segment, evaluated with a given traffic flow q, is computed as follows:

$$L_i = 42.2 + 10 \log_{10} q + C, \tag{12}$$

where C is the correction factor required for different values of speed v, percentage of heavy vehicles p and gradient of the road g. In fact, the basic computation of L_i (with $C = 0$) considers the average speed $v = 75$ Km/h, the percentage of heavy vehicles $p = 0\%$ and the gradient of the road $G = 0$ degrees.

In order to simulate scenarios different with respect to the basic one, a proper correction $C = C_1 + C_2$ needs to be applied. In particular, C_1 is the correction for v and p:

$$C_1 = 33 \log_{10} \left(v + 40 + \frac{500}{v} \right) + 10 \log_{10} \left(1 + \frac{5p}{v} \right) - 68.8 \tag{13}$$

while C_2 is the correction for the gradient of the road and is computed as:

$$C_2 = 0.3 \cdot g. \tag{14}$$

Finally, the contributions of the S segments are combined in order to calculate the overall traffic noise L_n:

$$L_n = 10 \log_{10} \sum_{i=1}^{S} 10^{L_i/10} \tag{15}$$

C. Discussion

The simulation has been performed by considering different scenarios our system can work on. In particular, we evaluate how the SNR varies depending on the following parameters: the distance d, the vehicle speed v in the set $\{50, 70, 100, 130\}$ Km/h, the number of vehicles per hour q in the set $\{100, 500, 1000, 4000\}$ vehicles/h.

In Table II we report the value of the parameters considered in the simulation, while the obtained results are reported in Fig. 4: in particular, each figure shows how the SNR (y-axis) varies with respect to the distance (x-axis) as the value of q is fixed. The curves on the same graphic refer to different values of v. As expected, it is evident how the SNR significantly decreases by increasing the speed, the traffic flow and the distance.

Although the considered model allows us to simulate the behavior of the proposed system in several environments by combining various traffic flows and vehicle speeds with different distances values, we decided to focus on the following two scenarios representing somehow the best and the worst case in which the proposed system can work: (1) a country road, where vehicles have typically a limited speed (around 50 Km/h) and the flow is very low (less than 100 vehicles/h); (2) a highway, where in the rush-hours the vehicle flow may be very high (around 4000 vehicles/h) as well as their speed (around 100 Km/h).

Taking into account, as we explain in detail in Section IV, that an event of interest with a $SNR = 10dB$ can be reliably detected by the proposed system, we designed the positioning of the microphones.

In Fig. 4a and Fig. 4d, we depict the attenuation of the SNR with respect to the distance at fixed traffic flow $q = 100$ and $q = 4000$, respectively. In the first case, we observe that the SNR of the sounds of interest is about $10dB$ at a distance of 120 meters, while in the second case a SNR of $10dB$ can be achieved at a distance of about 25 meters. This implies that for a country road the microphones can be placed at about 240 meters far from each other. The highway scenario, instead, is definitively more challenging due to the high number of vehicles crossing the road and the optimal distance between microphones is approximately $m = 50$ meters.

Parameter	Value	Description
L_s	140 dB	intensity level of the source
h_s	1 meter	height of the source
h_r	4 meter	height of the receiver
d_0	1 meter	reference distance
G	0	ground coefficient
α	32.8 dB/Km	atmospheric attenuation coefficient
p	5 %	percentage of heavy vehicles
g	3 %	gradient of the road
S	5	number of segments

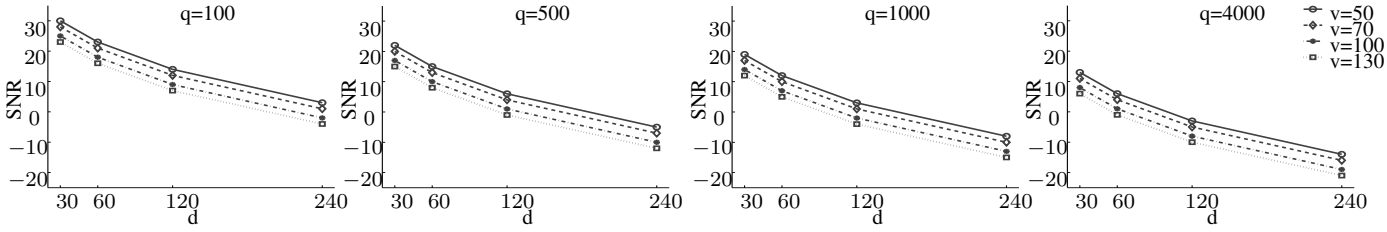
TABLE II: Summary of the values of the parameters used for the evaluation of the distance d.

IV. Experimental results

A. The data set

To the best of our knowledge, there are no publicly available data sets for road surveillance applications. Thus, we

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI 10.1109/TITS.2015.2470216, IEEE Transactions on Intelligent Transportation Systems



(a) Best case (b) Average case (c) Average case (d) Worst case

Fig. 4: Variation of the value of SNR (expressed in dB) with respect to the distance d (expressed in meters) the average speed v (expressed in Km/h) and the traffic flow q (expressed in vehicles/h).

created a data set that contains two classes of hazardous road events, namely crashes and tire skidding. The audio clips are sampled at 32 KHz, with a resolution of 16 bits per PCM sample; the whole data set was made publicly available at *http://mivia.unisa.it* for benchmarking purposes.

An audio-based system for road surveillance has to deal with different kinds of background sounds, ranging from very quiet background (i.e. in the country roads) to highly noisy traffic jams (i.e. in the center of a big city) and highways. Thus, in the proposed data set the events of interest are superimposed to different background sounds in order to simulate their occurrence in various environments. We, originally, collected 59 samples of crashes and 45 of tire skidding, together with the sound of 23 different road locations. We adopted a procedure to combine the original sounds, which we explain in the following.

The audio clips $x(n)$ have been, initially, normalized so that they have all the same overall energy:

$$\overline{x}(n) = \frac{x(n)}{x_{rms}(n)}. \tag{16}$$

where $x_{rms}(n)$ is the root mean square (RMS) value of the clip. A background clip $b(n)$ of about one minute duration is randomly selected from the typical traffic sounds. Then a number N_e of foreground events is randomly chosen from the original data set and superimposed to the background sound, in order to account for the occurrence of events of interest in a real environment. The selected events are mixed with the background sound, as follows:

$$out_j(n) = \sum_{i=1}^{N_e} \left\{ b_j(n) \oplus_{[s_i,e_i]} \left[A \cdot \overline{x}_i(n) \right] \right\}, \tag{17}$$

where $\oplus_{[s_i,e_i]}$ is an operator that combines the signal $\overline{x}_i(n)$ with the signal $b_j(n)$ in the interval delimited by $[s_i, e_i]$, starting and ending points of the sounds of interest, respectively. The point e_i is distanced from the starting point of the next sound s_{i+1} by an interval of 4 to 7 seconds in which only background sound is present. The attenuation (or amplification) factor A is determined so as to achieve a signal to noise ratio of $15dB$.

The final data set is composed of 57 audio clips of about one minute created with the procedure defined above. Each of the clips contains a sequence of events of interest: in total, 200 events per class are present. The produced clips are organized into $N = 4$ folds, each of them containing 50 events from each class of interest that overlap various traffic background sounds. The samples contained in a fold (both background and events of interest) are not present in the remaining folds, which are thus completely independent from each other. Moreover, high variability in the data is ensured by the heterogeneous background sounds on which the events of interest are superimposed. Within a given fold, the same event can be present as mixed with different backgrounds, in order to better represent various real situations. In the following of the text, we will refer to the different classes with the following abbreviations: *BN* for the background noise, *CC* for car crashes and *TS* for tire skidding. A detail of the composition of the data set is reported in Table III.

	#Events	Duration (s)
BN	-	2732
CC	200	326.3
TS	200	522.5

TABLE III: Details on the composition of the data set. The total duration of the sounds is expressed in seconds.

B. Experimental setup

For the computation of the low-level features, the audio stream is divided in frames of $T_f = 32$ milliseconds corresponding to 1024 PCM samples. We found that the choice of $T_f = 32ms$ is a reasonable compromise to take into account both low- and high-frequency properties of the signal and to perform a reliable short-time analysis of audio stream sampled ad $32KHz$. Two consecutive frames are overlapped for the 75% of their length in order to ensure continuity in the analysis of the audio stream. Different values of the number K of clusters (from 64 to 1024) have been considered for the experiments in order to evaluate the sensitivity of the system.

The high-level feature vector is computed for a time window of 3 seconds that shifts forward by 1 second. Two consecutive time windows, thus, overlap by two seconds. In this way, the continuity of analysis is ensured also at a time resolution of the order of the seconds: events that occur at the end of one window fall roughly in the middle of the next one.

For the experiments, the N-fold cross-validation is used. Cross-validation is a technique used for the assessment of the performance of a pattern recognition system and of its generalization capabilities to different data. It consists in the separation of a data set into a number of folds, which are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI 10.1109/TITS.2015.2470216, IEEE Transactions on Intelligent Transportation Systems

independent from each other in terms of samples. It means that the samples contained in one fold are not present in other folds. The cross-validation is often used to estimate how accurately a system will work in practice and how stable it will be under different conditions. In turn, $N-1$ folds are used as a training set to learn the classification model, and the remaining fold is used as a test set. The results of the N test obtained in this way are then averaged.

C. Performance evaluation

We evaluate the performance of the proposed system by measuring the recognition rate (true positive rate, TPR), i.e. the rate of correctly detected events of interest, and the false positive rate (FPR), i.e. the rate of wrongly detected events of interest when only background sound is present. A correct classification is counted when at least one of the overlapping time windows with the events is correctly classified. A false positive occurrence, which corresponds to a false alarm in a real system, is counted if an event of interest is detected when only background sound is present. In the case that the same event of interest is detected in two consecutive background time windows, only one false positive occurrence is counted.

Furthermore, we compute the receiver operating characteristic (ROC) curve, a method that is widely used to evaluate the overall performance of a classification system. It is a plot of the trade-off between the TPR and FPR of a classifier as its discrimination threshold is varied. The closer a ROC curve to the top-left corner of the plane, the better the performance. We consider the area under the ROC curve (AUC), which is equal to 1 for a perfect system, as an overall measure of the performance.

In Fig. 5 we report the performance of the proposed system (red solid line) in terms of recognition rate on the data set. We studied the variation of the recognition rate with respect to the number of basic audio words (clusters) learned during the training phase. In the top row of Fig. 5 the performance of the SVM classifier is depicted for the three considered sets of low-level features. We achieve an average recognition rate of 82%, 80.25% and 75% with a standard deviation of 1.5, 1.64 and 2.4 by employing as low-level features the set proposed in [24], the MFCC and the BARK, respectively. Moreover, we estimated the variance of the generalization error for the 4-folds cross-validation using the method of Nadeau and Bengio [49]. We observed that the estimated variance is from 25 to 50 times smaller that the average error, thus confirming statistical significance of the experiments on $N = 4$ folds.

In addition to the SVM classifier, we employed a k-Nearest Neighbor (kNN) classifier in order to evaluate the generalization capabilities of the proposed high-level representation. We depict the performance achieved with the kNN classifier in the bottom row of Fig. 5. The value of k has been experimentally set to 5. Although the performance results of the SVM-based classifier are stable with respect to the number of clusters, the performance achieved with the kNN classifier suggests that an increasing number of audio words causes a worsening of generalization capabilities. Thus, if too many words are used in the training phase, the system will be specialized in the recognition of the events from the training set. However, for the application at hand, the number of clusters is not a critical parameter as it is kept below 128.

Results on the data set

	Rec. Rate	Miss Rate	Error Rate	FPR
Bark	75%	21%	4%	10.96%
Mfcc	80.25%	19%	0.75%	5.48%
[24]	82%	17.75%	0.25%	2.85%

TABLE IV: Detailed results achieved by the proposed system configured with $K = 64$ basic audio words.

Bark

		Guessed		
		CC	TS	Miss
True	CC	86.0%	4.5%	9.5%
	TS	2.0%	64.00%	34%

MFCC

		Guessed		
		CC	TS	Miss
True	CC	89.5%	1.0%	9.5%
	TS	0.5%	71.0%	28.5%

[24]

		Guessed		
		CC	TS	Miss
True	CC	89.0%	0%	11.0%
	TS	0.5%	75.0%	24.5%

TABLE V: Classification matrices achieved by the proposed method on the data set with the three considered sets of low-level features.

In Table IV we report a summary of the results achieved by the system configured with $K = 64$ clusters, which is the value that gives the highest generalization[2]. In Table V, instead, we report the classification matrices achieved by the proposed system. We can note that the features proposed in [24] and the MFCC features show higher robustness to traffic noise with respect to Bark features. This determines that the system achieves a larger false positive rate when the Bark features set is used, due to the difficulties in differentiating the basic units of the sounds of interest from the background in very noisy traffic conditions. However, further studies on the temporal integration of basic audio units could improve the robustness to noise and the detection capabilities.

D. Sensitivity analysis

In a real environment, the sound source can be located at various distances from the microphone, resulting in the acquisition of signals with different intensity and signal to noise ratio. We performed a sensitivity analysis of the proposed system with respect to the signal intensity and the number of clusters. We decreased the intensity of the signal by $-3dB$

[2]With $K = 64$ clusters, the SVM classifiers learn for the classes BN, CC and TS the following average number of support vectors: $(60, 55, 50)$ for Bark features set, $(55, 70, 60)$ for MFCC features set and $(50, 60, 55)$ for the feature set in [24].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI 10.1109/TITS.2015.2470216, IEEE Transactions on Intelligent Transportation Systems

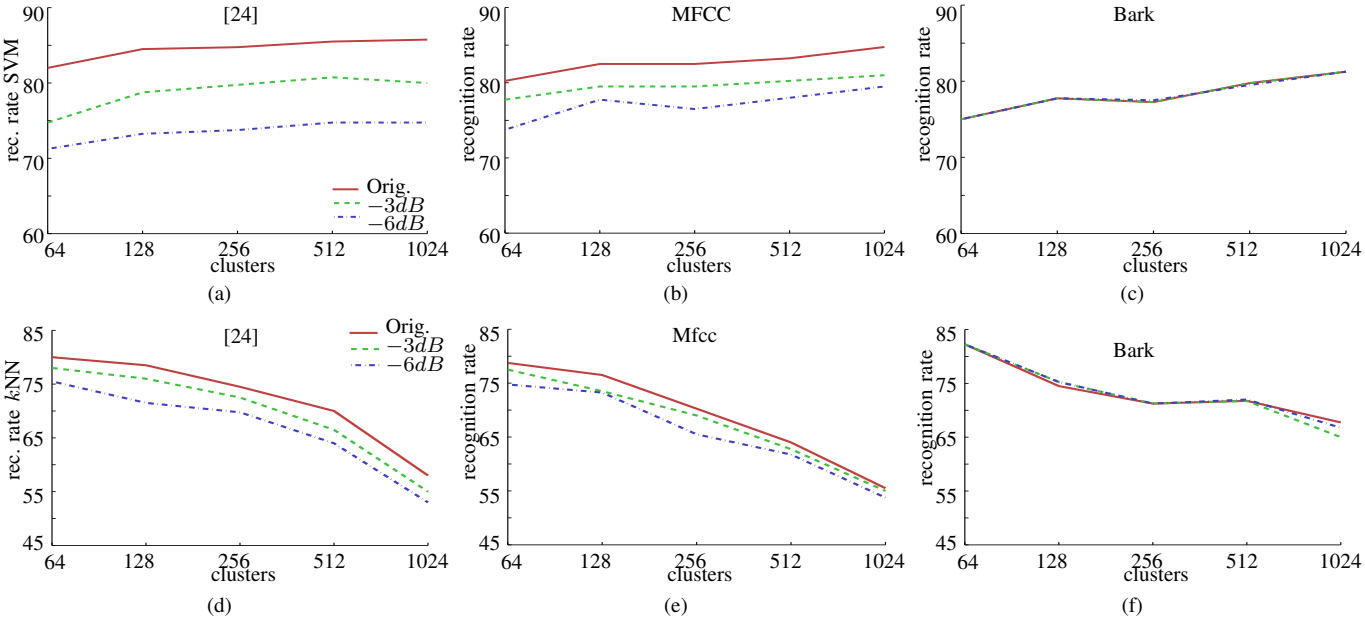


Fig. 5: On the first row the recognition rate of the proposed system (solid red line) for the three considered set of features is reported. The solid lines represent the performance obtained on a test set with the same SNR as the training set ($15dB$). The green and blue dashed lines show the results when the signal intensity (and the SNR) is reduced by $3dB$ and $6dB$, respectively. On the second row, the performance achieved with the kNN classifier demonstrate a loss in generalization capabilities of the proposed method when a too high number of basic audio words is chosen.

Sensitivity analysis

	[24]		MFCC		Bark	
	Rec.	σ	Rec.	σ	Rec.	σ
SVM (Orig.)	84.50%	1.50	82.65%	1.64	78.20%	2.40
SVM (Att.)	78.95%	4.94	79.80%	2.84	78.20%	2.20
kNN (Orig.)	72.20%	8.84	69%	9.49	77.30%	5.72
kNN (Att.)	69.52%	7.86	67.45%	8.45	77.30%	5.50

TABLE VI: Results of the performed sensitivity analysis. For both the employed classifiers the average recognition rate and its standard deviation are reported in the case of classification of the events in the proposed data set (orig.) and their attenuated version (Att.).

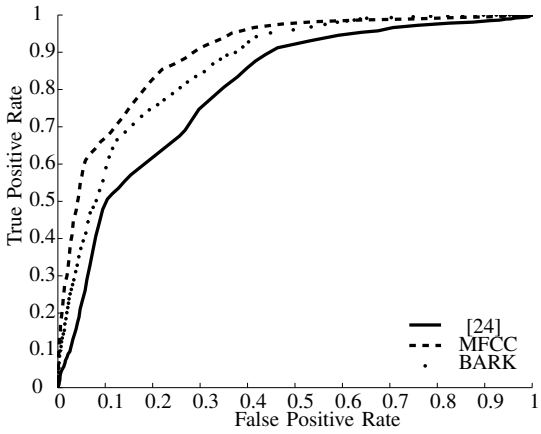


Fig. 6: ROC curves of the proposed system configured with the three considered sets of features.

and $-6dB$, in order to evaluate the detection capabilities at a distance of 25 and 120 meters depending on the scenario, according to the analysis presented in Section III. In practice, we trained the system on the events in the original data set and then tested it on events whose intensity is $-3dB$ and $-6dB$ of the original signal.

As observed in the previous paragraph, the number of basic audio words learned during the training process influences the generalization abilities of the system, while the trend of the recognition rate on the attenuated versions of the sounds (green and blue dashed lines for $-3dB$ and $-6dB$, respectively) is coherent with the one of the original data set.

Conversely, it is worth noting that the performance of the system with respect to different distances of the sound source depends mostly on the low-level representation of the audio signal. When temporal features based on the intensity and energy of the signal are used to describe the audio frames, in fact, the performance inevitably decreases with an increasing distance of the events from the microphone (blue and green lines in Fig. 5a, 5b, 5d and 5e). In such cases, when the energy of an event of interest decreases, it becomes comparable with the one of the background noise and it is more difficult to discriminate such events. The MFCC features, widely used for several audio recognition tasks like speech recognition or speaker identification, are sensitive to additive noise. However they show higher robustness to different signal to noise ratios, resulting in more stable results, as it can be seen in Fig. 6. From Fig. 5c and Fig. 5f, it is evident how the low-level features based on the distribution of the spectral energy in sub-bands has shown to be robust with respect to decreasing values of the power of the signal.

In Table VI we report the average recognition rate and its

standard deviation achieved by the proposed system varying the number of clusters for the test on the original data set and the one considering also the attenuated versions of the signals. The results registered by using the kNN classifier are highly influenced by the loss in generalization capabilities when a high number of cluster is configured. In Fig. 6, instead, we compare the ROC curves achieved by using the three sets of low-level features. The area under the curves (AUC) are equal to 0.80, 0.90 and 0.86 for the features used in [24], the MFCC and BARK, respectively. The ROC analysis confirms that the features based on the intensity and energy of the signal [24] are inadequate for recognition of sounds at various distances, while features based on frequency-analysis (MFCC and BARK) have higher robustness to different SNR.

E. Real-time performance

The algorithm utilizes about 3% of the resource of a single Intel i5 CPU core to process audio streams sampled at 32 KHz. It has been also implemented and runs in real-time on a STM32F4 board, making its deployment very inexpensive.

V. Conclusions

In this paper we proposed a system for detecting hazardous situations on roads by analyzing the audio stream acquired by surveillance microphones. We carried out the experiments on a data set that we created and made publicly available, with the aim of studying the sensitivity of the proposed system with respect to its configuration parameters. Furthermore, we conducted a careful design analysis in order to understand the potentiality of the proposed architecture, in terms of the maximum distance at which an event of interest can be still recognized in different kinds of environment, ranging from country roads to highways.

The achieved results confirm that the proposed system can be effectively used in noisy road environments, with an average accuracy of 78.95% at a maximum distance of 120 meters in country roads and of 25 meters on highways. Furthermore, its overall processing load is still compatible with low cost systems, so encouraging its porting on embedded systems with limited hardware resources. This property allows the realization of road surveillance systems with low deployment cost, also in combination with already existing surveillance architectures that provide audio acquisition sensors.

Acknowledgment

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

References

[1] T. Gandhi and M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 3, pp. 413–430, Sept 2007.

[2] S. Rauscher, G. Messner, and P. Baur, "Enhanced automatic collision notification system - improved rescue care due to injury prediction- first field experience," in *ESV*, 2009, pp. 1–10.

[3] J. White, C. Thompson, H. Turner, B. Dougherty, and D. Schmidt, "Wreckwatch: Automatic traffic accident detection and notification with smartphones," *Mobile Networks and Applications*, vol. 16, no. 3, pp. 285–303, 2011.

[4] S. Sivaraman, B. Morris, and M. Trivedi, "Observing on-road vehicle behavior: Issues, approaches, and perspectives," in *IEEE ITSC*, Oct 2013, pp. 1772–1777.

[5] L. Brun, A. Saggese, and M. Vento, "Dynamic scene understanding for behavior analysis based on string kernels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 10, pp. 1669–1681, Oct 2014.

[6] S. Sivaraman and M. Trivedi, "Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1773–1795, Dec 2013.

[7] L. Wang, N. Yung, and L. Xu, "Multiple-human tracking by iterative data association and detection update," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 1886–1899, Oct 2014.

[8] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. PP, no. 99, pp. 1–9, 2014.

[9] A. Abadi, T. Rajabioun, and P. Ioannou, "Traffic flow prediction for road transportation networks with limited traffic data," *IEEE Trans. Intell. Transp. Syst.*, vol. PP, no. 99, pp. 1–10, 2014.

[10] M. Cristani, M. Bicego, and V. Murino, "Audio-visual event recognition in surveillance video sequences," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 257–267, 2007.

[11] P. Marmaroli, M. Carmona, J.-M. Odobez, X. Falourd, and H. Lissek, "Observation of vehicle axles through pass-by noise: A strategy of microphone array design," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1654–1664, Dec 2013.

[12] M. A. Anusuya and S. K. Katti, "Speech recognition by machine, a review," *CoRR*, vol. abs/1001.2267, 2010.

[13] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Commun.*, vol. 56, no. 0, pp. 85–100, 2014.

[14] R. Cai, L. Lu, and A. Hanjalic, "Co-clustering for auditory scene categorization," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 596–606, 2008.

[15] H. Malik, "Acoustic environment identification and its applications to audio forensics," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 11, pp. 1827–1837, 2013.

[16] L. Cordella, P. Foggia, C. Sansone, and M. Vento, "A real-time text-independent speaker identification system," in *ICIAP*, 2003, pp. 632–637.

[17] Z. Saquib, N. Salam, R. Nair, N. Pandey, and A. Joshi, "A survey on automatic speaker recognition systems," in *Signal Processing and Multimedia*. Springer Berlin Heidelberg, 2010, vol. 123, pp. 134–145.

[18] V.-T. Vu, F. Bremond, G. Davini, M. Thonnat, Q.-C. Pham, N. Allezard, P. Sayd, J.-L. Rouas, S. Ambellouis, and A. Flancquart, "Audio-video event recognition system for public transport security," in *Crime and Security, 2006. The Institution of Engineering and Technology Conference on*, June 2006, pp. 414–419.

[19] W. Zajdel, J. Krijnders, T. Andringa, and D. Gavrila, "Cassandra: audio-video sensor fusion for aggression detection," in *IEEE AVSS*, Sept 2007, pp. 200–205.

[20] J.-L. Rouas, J. Louradour, and S. Ambellouis, "Audio events detection in public transport vehicle," in *IEEE ITSC*, 2006, pp. 733–738.

[21] P. Borkar and L. Malik, "Review on vehicular speed, density estimation and classification using acoustic signal," *Int. Journal for traffic and transport engineering*, 2013.

[22] S. Barnwal, R. Barnwal, R. Hegde, R. Singh, and B. Raj, "Doppler based speed estimation of vehicles using passive sensor," in *IEEE ICMEW*, July 2013, pp. 1–4.

[23] Q.-C. Pham, A. Lapeyronnie, C. Baudry, L. Lucat, P. Sayd, S. Ambellouis, D. Sodoyer, A. Flancquart, A.-C. Barcelo, F. Heer, F. Ganansia, and V. Delcourt, "Audio-video surveillance system for public transportation," in *IEEE IPTA*, July 2010, pp. 47–53.

[24] V. Carletti, P. Foggia, G. Percannella, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance using a bag of aural words classifier," in *IEEE AVSS*, Aug 2013, pp. 81–86.

[25] Z. Liu, Y. Wang, and T. Chen, "Audio Feature Extraction and Analysis for Scene Segmentation and Classification," *The Journal of VLSI Signal Processing*, vol. 20, no. 1, pp. 61–79, Oct. 1998.

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

[27] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of mfcc," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[28] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *The Journal of the Acoustical Society of America*, vol. 33, no. 2, pp. 248–248, 1961.

[29] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: a systematic review," *CoRR*, vol. abs/1409.7787, 2014.
[30] M. Vacher, D. Istrate, L. Besacier, J. F. Serignat, and E. Castelli, "Sound Detection and Classification for Medical Telesurvey," in *ICBME*, C. ACTA Press, Ed., Feb. 2004, pp. 395–398.
[31] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *ICME*, 2005, pp. 1306 –1309.
[32] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection in noisy environments," in *Proc. EURASIP European Signal Processing Conference*, Poznan, Poland, 2007.
[33] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *IEEE AVSS*, 2007, pp. 21–26.
[34] A. Rabaoui, M. Davy, S. Rossignol, and N. Ellouze, "Using one-class svms and wavelets for audio surveillance," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 4, pp. 763–775, 2008.
[35] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "Probabilistic novelty detection for acoustic surveillance under real-world conditions," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 713 –719, 2011.
[36] Q. Li, M. Zhang, and G. Xu, "A novel element detection method in audio sensor networks," *INT J DISTRIB SENS N*, 2013.
[37] P. Foggia, A. Saggese, N. Strisciuglio, and M. Vento, "Cascade classifiers trained on gammatonegrams for reliably detecting audio events," in *IEEE AVSS*, Aug 2014, pp. 50–55.
[38] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "An adaptive framework for acoustic monitoring of potential hazards," *EURASIP J. Audio Speech Music Process.*, vol. 2009, pp. 13:1–13:15, Jan. 2009.
[39] D. Conte, P. Foggia, G. Percannella, A. Saggese, and M. Vento, "An ensemble of rejecting classifiers for anomaly detection of audio events," in *IEEE AVSS*, Sept 2012, pp. 76–81.
[40] M. Chin and J. Burred, "Audio event detection based on layered symbolic sequence representations," in *IEEE ICASSP*, 2012, pp. 1953–1956.
[41] S. Technical Committee ISO/TC 43, Acoustics, "Iso 9613-2 - acoustics - attenuation of sound during propagation outdoors," 1996.
[42] C. Steele, "A critical review of some traffic noise prediction models," *Applied Acoustics*, vol. 62, no. 3, pp. 271 – 287, 2001.
[43] J. Quartieri, M. N., G. Iannone, C. Guarnaccia, D. S., T. A, and L. T., "A review of traffic noise predictive models," in *Recent Advances in Applied and Theoretical Mechanics*, 2009, pp. 72–80.
[44] U. K. D. of Environment and H. welsh Office Joint Publication, "Calculation of road traffic noise," 1975.
[45] B. für Verkehr, "Richtlinien für den lärmschutz an strassen," 1981.
[46] G. B. Canelli, K. Gluck, and S. S. A., "A mathematical model for evaluation and prediction of mean energy level of traffic noise in italian towns," *Acustica*, 1983.
[47] SETRA, CERTU, LCPC, and CSTB, "Bruit des infrastructures routieres : methode de calcul incluant les effets meteorologiques, version experimentale, nmpb-routes-96," 1995.
[48] N. P. Laboratory, "Technical guides - calculation of road traffic noise 1988," 2015. [Online]. Available: http://resource.npl.co.uk/acoustics/techguides/crtn/
[49] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Machine Learning*, vol. 52, no. 3, pp. 239–281, 2003.

Pasquale Foggia received the Laurea degree (cum laude) in computer engineering in 1995, and the Ph.D. degree in electronic and computer engineering in 1999, from the ÒFederico IIÓ University of Naples, Naples, Italy. He was an Associate Professor of computer science with the Department of Computer Science and Systems, University of Naples, from 2004 to 2008, and has been with the University of Salerno, Fisciano, Italy, since 2008. His current research interests include basic methodologies and applications in the fields of computer vision and pattern recognition. He is the author of several research papers on these subjects. Dr. Foggia has been a member of the International Association for Pattern Recognition (IAPR), and has been involved in the activities of the IAPR Technical Committee 15 (Graph-based Representations in Pattern Recognition) since 1997.

Nicolai Petkov received the Dr.sc.techn.degree in Computer Engineering (Informationstechnik) from Dresden University of Technology, Dresden, Germany. He is professor of computer science and head of the Intelligent Systems group of the Johann Bernoulli Institute of Mathematics and Computer Science of the University of Groningen, the Netherlands. He is the author of two monographs and coauthor of another book on parallel computing, holds four patents and has authored over 100 scientific papers. His current research is in image processing, computer vision and pattern recognition, and includes computer simulations of the visual system of the brain, brain-inspired computing, computer applications in health care and life sciences and creating computer programs for artistic expression. Prof. dr. Petkov is a member of the editorial boards of several journals

Alessia Saggese received in December 2010 the Laurea degree (cum laude) in Computer Engineering from the University of Salerno, Italy and in February 2014 the Ph.D. degree in electronic and computer engineering from the University of Salerno, Italy, and from the University of Caen Basse Normandie, France. She is currently an Assistant Professor of the University of Salerno. Since July 2012 she is a member of the IAPR Technical Committee 15 (Graph-based Representations in Pattern Recognition). Her research interests mainly pertain real time video analysis and interpretation for traffic monitoring and video surveillance applications.

Nicola Strisciuglio received the master degree (cum laude) in Computer Engineering from University of Salerno in 2012. Currently, he is a PhD student at the Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen (The Netherlands) and at Dept. of Information and Electrical Engineering and Applied Mathematics, University of Salerno (Italy). His current research interests include artificial intelligence, pattern recognition, machine learning, audio analysis and computer vision.

Mario Vento is a fellow scientist of the International Association Pattern Recognition (IAPR). Currently he is Full Professor of Computer Science and Artificial Intelligence at the University of Salerno (Italy), where he is the coordinator of the Artificial Vision Lab. From 2002 to 2006 he served as chairman of IAPR Technical Committee TC15 on Graph Based Representation in Pattern Recognition, and from 2003 as associate editor of the Electronic Letters on Computer Vision and Image Analysis. His research interests fall in the areas of Artificial Intelligence, Image Analysis, Pattern Recognition, Machine Learning and Computer Vision. More specifically, his research activity covered real time video analysis and interpretation for traffic monitoring and video surveillance applications, Classification Techniques, either Statistical, Syntactic and Structural, Exact and Inexact Graph Matching, Multi-Expert Classification and Learning Methodologies for Structural Descriptions. He authored over 200 research papers in International Journals and Conference Proceedings and serves as referee for many relevant journals in the field of Pattern Recognition and Machine Intelligence.