Audio surveillance

Stavros Ntalampiras
Artificial Intelligence Group, Electrical and Computer Engineering Department, University of Patras, Rio, Greece

Abstract

This chapter deals with a relatively new application domain of the generalized sound recognition technology, audio surveillance. The particular branch of computational auditory scene analysis aims at detecting acoustic events that may be indicative of catastrophic situations (e.g., gunshot, scream, etc.) in timely fashion. In general, this kind of systems is meant to help the authorized personnel through a decision support interface toward taking the appropriate actions for minimizing the effect of the hazard. This chapter provides a thorough analysis on the way that the generalized audio recognition technology can be adapted to the needs of audio surveillance. The acoustic parameters and the pattern recognition algorithms that can be used for the specific domain are explained. Subsequently, this work provides a representative picture of the bibliography and discusses several aspects that could be of interest with respect to future directions. Lastly, it mentions several privacy concerns along with conclusions, where the merits of surveillance frameworks that are based on heterogeneous modalities are emphasized.

Keywords: Computational Auditory Scene Analysis, Sound Event Detection, Audio Pattern Recognition, Civil Safety

1 Introduction

Nowadays surveillance is becoming a common practice in various environments, like stores, agencies, and so on. Detection of situations that may include any type of danger (human injuries, damage of properties, etc.) is of particular importance for civil safety. As a result, there is a need for unattended space monitoring, which has motivated the signal processing community toward experimenting with various frameworks. Surveillance systems are typically based on the visual modality since the information they capture may provide an accurate picture of the region of interest [1]. However, there are several problems that need to be handled, like the field of view of the sensor network for capturing the entire region as well as the fact that several scenes may look normal even though an

WIT Transactions on State of the Art in Science and Engineering, Vol 54, © 2012 WIT Press
www.witpress.com, ISSN 1755-8336 (on-line)
doi:10.2495/978-1-84564-562-5/12

atypical situation is in progress. On top of that the acoustic modality can capture information that may be difficult or even impossible to obtain by any other means. The basic advantages of the acoustic sensors over the visual ones are (a) lower computational needs during information processing and (b) the illumination conditions of the space to be monitored and/or possible occlusions do not have an immediate effect on sound.

In this chapter, *audio surveillance* includes capturing the audio information of a particular space and processing the incoming sequence toward detecting sound events that are indicative of catastrophic situations, that is, atypical sound events, for example, scream, gunshot, explosion, and so on. This definition clearly states that audio surveillance primarily constitutes a branch of the generalized sound recognition technology. The particular technology is a part of the scientific domain, which is often called computational auditory scene analysis (CASA), and aims at a complete description of the region of interest based solely on the acoustic modality. A complete description typically includes localization, enumeration, separation, and recognition of all the included acoustic emissions. Sound recognition has many interesting applications, which can be categorized as follows:

- Voice activity detection (VAD): The principal goal of a VAD algorithm is to segment an audio signal into speech and nonspeech parts. This process is to assist a speech/speaker recognition system by elaborating on speech segments alone, thus improving its performance.
- Applications as regards to processing of musical signals: Over the past decade, this application category has attracted the interest of a relatively large number of researchers [2–4]. It includes applications such as music transcription, identification of music genre, recognition of performer, indexing and retrieval of musical data, and so on.
- Applications as regards to processing of bioacoustic signals: This special kind of audio signals belongs to very different frequency ranges. Animal vocalizations may be employed for mate attraction, territorial defense, and so on. There exists a variety of applications, like tracking of animals, monitoring of endangered species, biodiversity indexing [5–7], and so on.
- Applications of machine acoustics signal processing: This area encompasses processing of acoustic signals emitted by solids (e.g., metal, rock, ceramic, etc.) when they are subjected to stress. These emissions can be characteristic of internal fracture and/or deformation. The associated applications are nondestructive testing, fault detection and function control, maintenance services [8,9], and so on.
- Context recognition: The specific application domain essentially comprises the recognition of the physical environment around a device, including identification of relevant sound events as well as recognition of the activity of the user. Context recognition gives the ability to a device to alter its functions according to the surrounding environment [10]. Other applications are memory extension [11], environment recognition for robots [12], acoustic surveillance [13], and so on.

This chapter is organized as follows: Section 2 focuses on the domain of generalized sound recognition as seen from the scope of audio surveillance. Section 3 provides an overview of the literature along with evaluation methodologies that are usually employed. Subsequently, Section 4 mentions some privacy concerns, while conclusions are drawn in Section 5.

2 Sound recognition for audio surveillance

The domain of audio recognition is currently dominated by techniques that are mainly applied to speech technology [14]. This fact is based on the assumption that all audio streams can be processed in a common manner, even if they are emitted by different sources. In general, the goal of generalized audio recognition technology is the construction of a system that can efficiently recognize its surrounding environment by solely exploiting the acoustic modality. Every sound source exhibits a consistent acoustic pattern that results in a specific way of distributing its energy on its frequency content. This unique pattern can be discovered and modeled by using statistical pattern recognition algorithms. Similarly, an audio surveillance system models and subsequently identifies the spectral patterns of atypical sound events. However, there exists a variety of obstacles that need to be tackled when such a system operates under real-world conditions. When we have to deal with a large number of different sound classes, the recognition performance is decreased. Moreover, the categorization of sounds into distinct classes is sometimes ambiguous (an audio category may overlap with another), while composite real-world sound scenes can be very difficult to analyze. This fact has led to solutions that target specific problems, while a generic system is still an open research subject.

A typical sound recognition system as regards to classification of N sound categories is depicted in Figure 1. Initially, the audio signal passes through a preprocessing step, which usually includes mean value removal and gain normalization. This stage is to remove inconsistencies for facilitating the parameterization step. Preprocessing is of particular importance with respect to acoustic surveillance toward avoiding any loss of information. DC offset appears in the case where a waveform has unequal quantities in the positive and negative spaces. Our scope is the signal to have its middle point at zero for obtaining the maximum dynamic range. Furthermore, it is usual for an abnormal sound event to demonstrate the “clipping” effect. Gain normalization scales the audio data so that the amplitude of the respective waveform is increased to the maximum level without introducing any type of distortion. Subsequently, the signal is segmented into frames of predefined size using a windowing technique (e.g., Hamming). Then the hypothesis is made that inside a particular frame the characteristics of the audio signal are stationary. Moreover, an overlap is usually inserted with respect to adjacent frames for smoothing any discontinuities. Various frame and overlap sizes have been reported in the literature (30–200 ms). The optimal choice depends on the specifics of the particular application, while it should

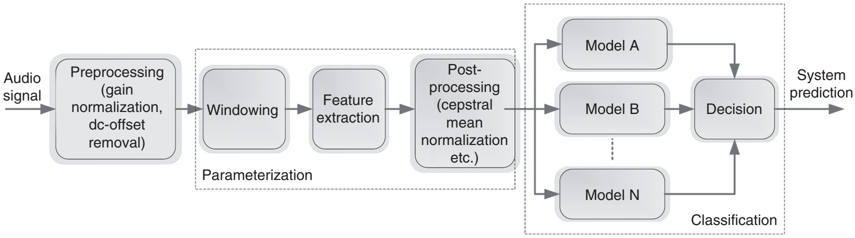

Figure 1: A sound recognition system as regards to classification of *N* sound categories.

be made after extensive experimentations. Afterwards, a feature extraction methodology is applied onto each frame. Feature extraction is a data reduction procedure, while its purpose is to summarize the audio segments using low-dimensional vectors. These vectors should capture the most relevant information with respect to a specific classification task. It should be noted that the inclusion of nonrelevant information may result in decreased performance. For example, in the audio surveillance case when one has to classify between explosions and screams, one should use features that are able to capture the differences between these two types of signals, for example, Mel frequency cepstral coefficients (MFCC) and features based on the Teager energy operator [15]. The following feature sets, which are typically used for sound recognition, can be employed for the special case of audio surveillance:

- MFCC: They originate from the speech/speaker recognition field. Their basic purpose is to mimic the human auditory system to some extent. More specifically, during their computation the nonlinearity of pitch perception as well as the nonlinear relationship between intensity and loudness are considered. In combination with their low computational cost, they have become the standard choice for many speech-related tasks, such as language identification, emotion recognition, and so on.
- The block diagram with respect to MFCC's extraction is depicted in Figure 2. Initially the signal is cut into frames of small duration (20–40 ms) based on the Hamming window technique. At this stage, a hop-size of 10 ms is usually employed. Afterwards the short-time Discrete Fourier Transform (DFT) is calculated for each frame using a predefined number of points (e.g., 256 or 512). A triangular filter bank elaborates on the outcome of the DFT. Subsequently, the data are logarithmically spaced and the Discrete Cosine Transform (DCT) is applied for exploiting its energy compaction properties.
- MPEG-7 low-level descriptors (LLD) [16]: MPEG-7 provides a set of standardized tools for automatic multimedia content description and offers a degree of "explanation" of the information meaning. It eases navigation of audio data by providing a general framework for efficient audio management. Furthermore, it includes a group of fundamental descriptors and description schemes

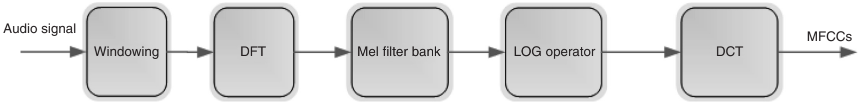

Figure 2: Extraction of MFCC.

for indexing and retrieval of audio data. Seventeen temporal and spectral descriptors that are useful for generalized sound recognition are used within the MPEG-7 audio standard. Several of them are quite simplistic (e.g., Audio Power) while others mainly target music processing (e.g., the ones that belong to the timbral group). The LLDs that may be proven effective as regards to the task of audio surveillance are as follows:

a) *Audio spectrum envelope*: This series of features belong to the basic spectral descriptors and is derived for the generation of a reduced spectrogram of the original audio signal. It is a log-frequency power spectrum and calculated by summing the energy of the original power spectrum within a series of logarithmically distributed frequency bands using a predefined resolution.
b) *Audio spectrum centroid*: The center of the log-frequency spectrum's gravity is given by this descriptor. Omitting power coefficients bellow 62.5 Hz (which are represented by a single coefficient) makes able the avoidance of the effect of a nonzero DC component.
c) *Audio spectrum spread*: The specific LLD is a measure of signal's spectral shape and corresponds to the second central moment of the log-frequency spectrum. It is computed by taking the root mean square deviation of the spectrum from its centroid.
d) *Audio spectrum flatness*: This descriptor is a measure of how flat a particular portion of the spectrum of the signal is and represents the deviation of the signal's power spectrum from a flat shape. The power coefficients are taken from nonoverlapping frames, while the spectrum is typically divided into ¼-octave resolution logarithmically spaced overlapping frequency bands. The ASF is derived as the ratio of the geometric mean and the arithmetic mean of the spectral power coefficients within a band.

The next stage of the sound recognition methodology, which is illustrated in Figure 1, is the post-processing of the extracted features. Techniques for normalizing the cepstral coefficients and/or the dynamic range are usually included. Both can be proven helpful for acoustic surveillance since they may reduce the dereverberation effects that usually appear when it comes to real-world conditions. Moreover, they can help during the classification stage, since they allow for a better comparison of the underlying characteristics that are exhibited by the novel data with the ones of the training data.

Another type of post-processing that can be used concurrently targets at projecting the feature coefficients onto a low-dimensional space. These processes try to keep only a small amount of the feature coefficients, which include their most important information. Even though feature projection facilitates the classification stage (since high-dimensional data tend to lead to a sparse representation), one should take extra care in order not to discard important information. The dimensionality reduction techniques that are proposed by the MPEG-7 audio standard are singular value decomposition, principal component analysis, nonnegative factorization, and independent component analysis. These can be used for audio surveillance tasks while keeping in mind that their majority are data depended approaches, which means that a large deviation between the train and test data may lead to disappointing recognition accuracy.

As a general comment on signal parameterization with respect to the area of audio surveillance, we claim that the features that provide a description of the spectrum are the most useful ones since the Fourier transform can efficiently characterize pressure waves. We believe that the MFCCs can provide a strong basis for the formulation of an effective feature set. MPEG-7 LLDs that characterize the signal in a different way can be appended. Their selection depends on the needs of the specific application. In addition, descriptors derived from the wavelet domain can also be used since their combination with the spectral ones have been shown to lead to improved recognition accuracy as regards to generalized sound classification [17]. The particular domain has not been fully explored as regards to atypical sound event detection, and we think that it could be very interesting for future research. Finally, our suggestion is to employ the DCT for the post-processing of the final feature vector since it is almost as efficient as the data-driven approaches at a much lower computational cost.

The final step of Figure 1 is the classification. The classifiers that are currently employed by the audio recognition community can be divided into two categories: discriminative and nondiscriminative. The discriminative ones try to approximate a boundary between the categories of the training data. Some examples are the polynomial classifier [18], multilayer perceptron [19], and Support Vector Machines (SVMs) [20]. On the opposite side, the generative approaches, which are the main class of the nondiscriminative classifiers, try to estimate the underlying distribution of the training data. They include Gaussian mixture models (GMMs) [21], hidden Markov models (HMMs) [22], and probabilistic neural networks (PNN) [23]. Other nondiscriminative approaches are the *k*-nearest neighbors (*k*-NN) [24] and the learning vector quantization (LVQ) [25]. In addition, several hybrid classification schemes have been reported in the literature [26–28], which exploit the merits of the two types of classification approaches. The majority of the audio surveillance frameworks that exist in the literature are based on generative approaches since these approaches tend to provide high recognition rates. However, there is still room for improvement and the most promising way to achieve higher performance is the development of hybrid methods. This kind of methods can be adjusted so as to satisfy the requirements of a specific application and potentially provide improved results.

3 A representative picture of the related literature

This section intents to provide a representative picture of what has been developed so far in the area of audio surveillance (see also Table 1). The emphasis of previous approaches is mainly placed on the classifier, the feature extraction process, the training data, and the number of classes. The system of Ntalampiras *et al.* [29] exploits the advantages of maximum a posteriori adaptation as well as diverse feature sets that allow detection of scream, normal speech, background environment, gunshot, and explosion sound events. The authors report results after a continuous operation for three subsequent days while using three types of environmental noise (metro station, urban and military environment). Their database was formed by using a combination of professional sound effect collections. Valenzise *et al.* [30] presented a surveillance system for gunshot and scream detection and localization in a public square. Forty-nine features were computed in total and given as an input to a hybrid filter/wrapper selection method. Its output was used to build two parallel GMMs for identifying screams from noise and gunshots from noise. Data were drawn out from movie sound tracks, Internet repositories, and people shouting at a microphone while the noise samples were captured in a public square of Milan. An interesting application, crime detection inside elevators, was explained in [31]. Their approach relied on time-series analysis and signal segmentation. Consistent patterns were discovered and the respective data were used for training one GMM for each one of the eight classes using low-level features. The data set contained recordings of suspicious activities in elevators and some event-free clips while they reported detection of all the suspicious activities without any misses. A gunshot detection method under noisy environments was explained in [32]. Their corpus consisted of data that were artificially created from a set of multiple public places and gunshot sound events extracted from the national French public radio. Widely used features were employed, including MFCC for constructing two GMMs with respect to gunshot and normal class using data of various Signal-to-Noise Ratio (SNR) levels. In [33] the issue of detection of audio events in public transport vehicles was addressed by using both a generative and a discriminative method. The audio data were recorded using four microphones during four different scenarios, which included fight scenes, a violent robbery scene, and scenes of bag or mobile snatching. They used GMM and SVM while their feature set was formed from the first 12 MFCC, energy, derivatives, and accelerations. Vacher *et al.* [34] presented a framework for sound detection and classification for medical telesurvey. Their corpus consisted of recordings made in the CLIPS laboratory, files of the "Sound Scene Database in Real Acoustical Environment" (Real World Computing Partnership* (RCWP) Japan). They used wavelet-based cepstral coefficients to train GMMs for eight sound classes while their system was evaluated under different SNR conditions. A hierarchical classification scheme that identified

* http://tosa.mri.co.jp/sounddb/indexe.htm

Table 1: Various approaches on the task of acoustic surveillance.

Reference	Atypical sound classes	Model adaptation	Classifier	Features	Database
Ntalampiras *et al.* [29]	Scream, gunshot, and explosion	MAP adaptation of GMMs	GMM	MFCC, MPEG-7, CB-TEO, Intonation	Large audio corpora from professional sound effects collections
Valenzise *et al.* [30]	Scream and gunshot	–	GMM	Temporal, spectral, cepstral, correlation	Movie soundtracks, Internet, and people shouts
Radhakrishnan & Divakaran [31]	Banging and non-neutral speech	–	GMM	MFCC	Elevator recordings
Clavel *et al.* [32]	Gunshot	–	GMM	MFCC, spectral moments	CDs for the national French public radio
Rouas *et al.* [33]	Shout	Adaptive threshold for sound activity detection	GMM, SVM	Energy, MFCC	Recorded during four scenarios
Vacher *et al.* [34]	Scream and glass break	–	GMM	Wavelet-based cepstral coefficients	Laboratory recordings and RCWP
Atrey *et al.* [35]	Shout	–	GMM	ZCR, LPC, LPCC, LFCC	Recorded in office corridor
Ito *et al.* [36]	Glass clash, scream, fire cracker	Adaptive threshold for abnormal sound event detection	GMM	MFCC, Power	Recorded under laboratory conditions

normal from excited sound events was described in [35]. The authors used four audio features for training GMMs, each one associated with one node of the classification tree. The audio was recorded for around two hours in the real environment (office corridor) and included talk, shout, knock, and footsteps. In [36] the authors use a multistage schema based on GMMs that “learns” the normal sounds and subsequently detects events that exhibit large differences from the normal ones. A procedure for automatic determination of the threshold that differentiates between normal and abnormal sounds is also reported. Their feature vector includes the 16 first MFCCs and the power along with the corresponding derivatives while their experiments took place on recordings made under laboratory conditions.

It is argued that previous research in the specific domain is far from concluding on a common framework as, for example, in the case of speech/speaker recognition where the classifier and the feature extraction process is more or less established (i.e., GMMs and HMMs as classifiers and variations of spectral features as input). The difficulty basically lies on the next three facts:

1. An atypical situation is not a well-defined category (e.g., laughter vs cry vs scream).
2. There are many cases where there is a thin line between a typical and an atypical situation (e.g., gunshot vs explosion).
3. The microphone can be located far from the source of the acoustic incident; therefore, reverberation and acoustic events belonging to an almost unrestricted range of classes may become the input to the microphone.

As a general conclusion, we can point out the fact that statistical-based approaches are used by the majority of the authors, while for each article the feature set is chosen so as to fit the needs of the specific application. An interesting direction to follow would be the establishment of frameworks that include hybrid methods during the pattern recognition phase, such as the combination of generative and discriminative approaches. Furthermore, because of the unavailability of real-world atypical audio data that include extreme emotional manifestations and abnormal sound events, the novelty detection methodology [37], which is only partially explored in [36], could be proven useful.

3.1 Evaluation of audio surveillance frameworks

The present section comments on a highly important issue of audio surveillance frameworks, that is, the evaluation methodology. The establishment of a common evaluation metric is critical toward making a reliable comparison between different surveillance approaches. This kind of frameworks essentially includes single or multiple detection problems. A typical representation technique of the performance of a detector is the receiver operating characteristic curve (ROC curve). An example of a ROC curve is illustrated in Figure 3. In this case, the true-positive rate ($R_{TP} = TP/(FN + TP)$, percentage of the correct classified test

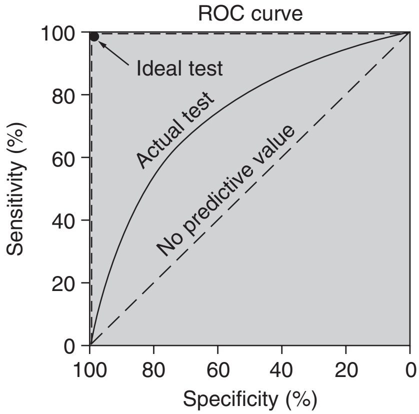

Figure 3: The ROC curve. The more the curve is in the upper left corner, the better is the detection system. The threshold T is increased from left to right. If the threshold is zero, every test case is classified as TRUE, thus the Sensitivity is one but the Specificity is zero (lower left corner); if the threshold is maximal, all test cases are classified as FALSE, thus the curve ends in the upper right corner. The optimal value for a threshold T is the one for which the curve is next to the upper left corner. The dashed line indicates the performance of a system that just is guessing (50% detection rate in a two class-problem). Source: Image Characteristics and quality, Terry Sprawls, www.sprawls.org.

cases from all of those that are "positive" in reality) is plotted in relation to the false-positive rate ($R_{FP} = FP/(FP + TN)$, percentage of test cases that are "negative" in reality and wrongly classified as "positive" by the detector) in dependence of a parameter, typically a threshold. This is done if an adjustable threshold T in the detection system is responsible for the decision "detected" or "not detected", and the optimal value for this threshold, a maximal quotient, should be found.

Although this type of error analysis provides useful information, it is believed that the Detection Error Tradeoff (DET) curves that comprise an adapted version of ROC curves should be used. A typical DET curve is depicted in Figure 4. The DET curves as introduced by the National Institute of Standards and Technology [38] can be viewed as presenting the trade-off between two error types: missed detections and false alarms. The point where the average of the missed detection and false alarm rates is minimized is the optimal point, that is, the one that should be used during the operation of the system. The specific average essentially is the cost function of a DET curve. There are two important things to note about the DET curve. First, in the case that the resulting curves are straight lines, it can safely be assumed that the underlying likelihood distributions from the system are normal. Second, the diagonal $y = -x$ on the normal deviate scale represents random performance. With a large number of targets and roughly equal occurrences of all nontargets, the overall performance is effectively represented.

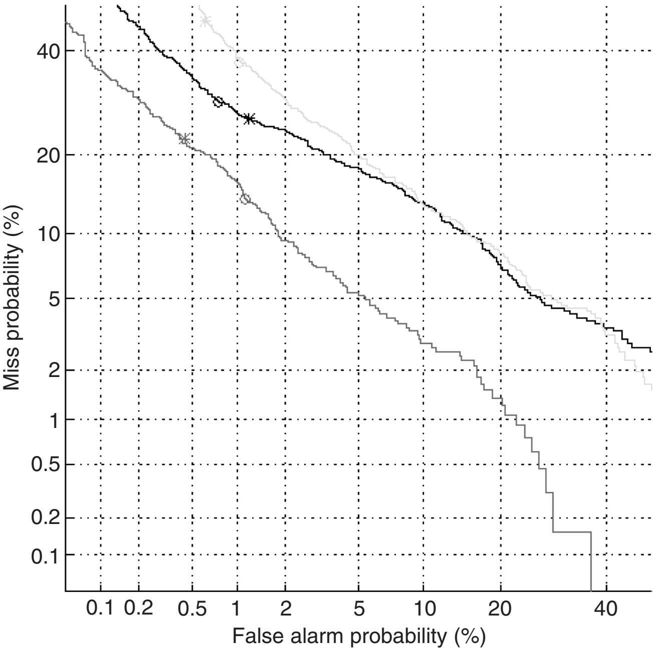

Figure 4: The DET curves can be viewed as presenting the trade-off between two error types: missed detections and false alarms. The "circles" on the curves indicate the optimal decision cost for a system, and the "starts" indicate the actual decision cost at the selected operational point (threshold, etc.). Source: Reference [38].

Unlike the standard ROC curve, the DET curves are approximately linear curves that are easily observable and suitable for presenting performance results where trade-offs of two error types are involved. Furthermore, the production of a DET curve requires a common scale for the likelihood of each event, which is a desirable property in many applications. Finally, the DET curve may include a number of special points to facilitate performance analysis, such as a specific false alarm or a miss detection rate.

4 Privacy

Acoustic sensors are sometimes perceived as invasive, especially when the subject of attendance is human. Therefore, privacy issues need to be fully taken into account while constructing an audio surveillance system. The research conducted in the specific scientific area is based and motivated by the next presuppositions:

- Threatening situations such as crime and terrorist acts in large urban areas are not fictitious scenarios but real facts that require special attention and measures.

Moreover, the knowledge that public spaces are being secured by intelligent monitoring is expected to discourage the manifestation of such acts.

- Surveillance, in general, is not in conflict with the law, and it is common practice in stores, agencies, airports, and so on, where the need for increased security justifies the installation of video cameras.
- Unattended autonomous surveillance is much less “invasive” as it precludes human interference from the interpretation of the sensor’s information as well as data broadcasting at any stage of the inference process. Therefore, it restricts human processing as well as unscrupulous circulation of personal data. As the right to privacy is claimed by more and more people, unattended surveillance ensures that the interpretation of the sensors information does not involve unauthorized human interference at any stage of the inference process.
- The main task of unattended surveillance is to identify in time the sensed situation and deliver the necessary warning messages to an authorized officer. It does not involve any other kind of uncontrolled action or initiative in part of the machine. In addition, the microphones are not used to identify individuals or to interpret spoken words or sentences.

It is believed that compliance with the above four points ensure that the privacy of all individuals is not to be compromised at any stage of the processing chain of an audio surveillance framework. Therefore, such frameworks can play a significant role with respect to civil safety [39].

5 Conclusion

Unattended space monitoring based on the acoustic modality comprises an effective tool toward scene analysis for detection of catastrophic situations. This chapter provided an overview of the technology that lies behind the specific scientific area, a descriptive review of the literature along with several privacy issues that need special attention. Although it is not possible to identify a general purpose feature set as well as the recognition technique that performs best for all surveillance applications, the usage of MFCCs as the starting point is suggested. MPEG-7 LLDs as well as other application-specific parameters can be appended at a second stage for improving classification accuracy. With respect to the pattern recognition part, the HMM approach is a reasonable choice since they offer satisfying detection results in many audio classification applications. Furthermore, a synthetic scheme that employs the complementary properties of the generative (e.g., GMM) and discriminative (e.g., SVM) classifiers can be employed.

Throughout this chapter, some directions for future research were suggested, which concentrate on acoustic signal processing. Another interesting direction to be explored is the combination of the acoustic sensors with other heterogeneous ones. The acoustic modality can play either a stand-alone role or be used in parallel with other modalities toward obtaining an enhanced analysis of the scene of interest. The information from heterogeneous sensors, which is both complementary and redundant, aims at surpassing the weaknesses of each modality in dealing with coverage of the sensed area and its response to occlusion, noise, and differing environmental conditions. These sensors can be complementary in two different ways:

1. The combination of different sensors' reports can be merged into a single but more complete piece of information.
2. Information gained from one sensing modality can be used to validate observations and/or aid the processing chain of the others.

For example, a network of proximity indicators can efficiently detect and count the number of people. However, this type of sensor gives no information about the height of the people involved or their appearance. Estimates of height, color of clothes, and appearance can be generated using observations from a monocular camera. However, if there are shadows or low/time varying lighting conditions or occlusion due to another person, the camera (which also generates ambiguity due to depth) will sense reflected light so that the image will be a product of both intrinsic skin reflectivity and external incident illumination and will, therefore, return poor results. Moreover, variations in ambient illumination will enlarge the within-class variability of any statistical classifier, thus severely degrading classification performance of subjects, behavior, and interaction. The detection of human presence and the complementary data of height but not of color can be provided by the infrared camera that detects the thermal emission of bodies (which is an intrinsic measurement that can be isolated from external illumination) and, therefore, works under low-light conditions. Acoustic data picked up by a microphone array and their associated time-frequency signatures can return bearing and location measurements as well as provide information for scene interpretation. To summarize, a dispersed network of multimodal sensors allows complementary views about the state of the environment to be deduced that would be unavailable to either sensor working alone.

References

[1] Haritaoglu, I., Harwood, D., & Davis, L., W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22(8)**, pp. 809–830, 2000.

[2] Eronen, A., & Klapuri, A., Musical instrument recognition using cepstral coefficients and temporal features. Proceedings of the ICASSP'00, Istanbul, pp. 753–756, 2000.

[3] Tzanetakis, G. & Cook, P., Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, **10(5)**, pp. 293–302, 2002.

[4] FitzGerald, D., Coyle, E., & Lawlor, B., Sub-band independent subspace analysis for drum transcription. Proceedings of the DAFX'02, Hamburg, pp. 65–69, September 2002.

[5] Ashiya, T., Hagiwara, M., & Nakagawa, M., IOSES: An indoor observation system based on environmental sounds recognition using a neural network. *Transactions of the Institute of Electrical Engineers of Japan*, **116-C(3)**, pp. 341–349, 1996.

[6] Gillespie, D. & Chappell, O., An automatic system for detecting and classifying the vocalizations of harbour porpoises. *Bioacoustics*, **13**, 37–61, 2002.

[7] Hennig, R.M., Acoustic feature extraction by cross-correlation in crickets. *Journal of comparative physiology. A Neuroethology, sensory, neural, and behavioral physiology*, **189**, pp. 589–598, 2003.

[8] Dimla, D.E., Jr, Lister, P.M., & Leighton, N.J., Neural network solutions to the tool condition monitoring problem in metal cutting. A critical review of methods. *International Journal of Machine Tools Manufacturing*, **37(9)**, pp. 1219–1240, 1997.

[9] Diei, E.N. & Dornfeld, D.A., Acoustic emission sensing of tool wear in face milling. *Transactions of ASME, Journal of Engineering for Industry*, **109**, pp. 234–240, 1987.

[10] Peltonen, V., Computational auditory scene recognition. Master of Science Thesis, Department of Information Technology, Tampere University of Technology, Tampere, 2001.

[11] Vemuri, S., Schmandt, C., Bender, W., Tellex, S., & Lassey, B., An audio-based personal memory aid. Proceedings of the 6th International Conference Ubiquitous Computing, Ubicomp'04, Tokyo, pp. 400–417, 2004.

[12] Chu, S., Narayanan, S., Jay Kuo, C.-C., & Matarić, M.J., Where am I? Scene recognition for mobile robots using audio features. Proceedings of the ICME'06, Ischia, pp. 885–888, 2006.

[13] Ntalampiras, S., Potamitis, I., & Fakotakis, N., On acoustic surveillance of hazardous situations. Proceedings of the ICASSP '09, Taipei, pp. 165–168, 2009.

[14] Foote, J.T., An overview of audio information retrieval. *ACM-Springer Multimedia Systems*, **7(1)**, pp. 2–11, 1999.

[15] Zhoun, G., Hansen, J.H.L., & Kaiser, J.F., Nonlinear feature based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing*, **9(3)**, pp. 201–216, 2001.

[16] Casey, M., MPEG-7 sound recognition tools. *IEEE Transactions on Circuits and Systems for Video Technology*, **11(6)**, pp. 737–747, 2001.

[17] Ntalampiras, S., Potamitis, I., & Fakotakis, N., Exploiting temporal feature integration for generalized sound recognition. *EURASIP Journal on Advances in Signal Processing*, article ID: 807162, **2009**, 2009.

[18] Specht, D.F., Generation of polynomial discriminant functions for pattern recognition. *IEEE Transactions on Electronic Computers*, **16**, pp. 308–319, 1967.

[19] Rosenblatt, F., The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**, pp. 386–408, 1958.

[20] Vapnik, V.N., *The Nature of Statistical Learning Theory*, New York: Springer, 1995.

[21] Peltonen, V., Tuomi, J., Klapuri, A., Huopaniemi, J., & Sorsa, T., Computational auditory scene recognition. Proceedings of the ICASSP'02, Orlando, pp. 1941–1944, May 2002.

[22] Casey, M., General sound classification and similarity in MPEG-7. *Organised Sound*, **6(2)**, pp. 153–164, 2001.

[23] Bolat, B. & Kucuk, U., Musical sound recognition by active learning PNN. *Lecture Notes in Computer Science, vol. 4105/2006, Multimedia Content Representation, Classification and Security*, ISSN:0302-9743, Springer Berlin/Heidelberg, pp. 474–481, 2006.

[24] Essid, S., Classification of audio signals: Machine recognition of musical instruments. Seminars, CNRS-LTCI, 2006.

[25] Yella, S., Gupta, N.K., & Dougherty, M., Pattern recognition approach for the automatic classification of data from impact acoustics. Proceedings of the AISC'2006, Palma De Mallorca, pp. 144–149, 2006.

[26] Dietterich, T., An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, **40(2)**, pp. 139–157, 2000.

[27] Kittler, J., Hatef, M., Duin, R., & Matas, J., On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20(3)**, pp. 226–239, 1998.

[28] Alkoot, F.M. & Kittler, J., Experimental evaluation of expert fusion strategies. *Pattern Recognition Letters*, **20(11)**, pp. 11–13, 1999.

[29] Ntalampiras, S., Potamitis, I., & Fakotakis, N., An adaptive framework for acoustic monitoring of potential hazards. *EURASIP Journal on Audio, Speech and Music Processing*, article ID: 594103, **2009**, 2009.

[30] Valenzise, G., Gerosa, L., Tagliasacchi, M., Antonacci, F., & Sarti, A., Scream and gunshot detection and localization for audio-surveillance systems. Proceedings of the Advanced Video and Signal-Based Surveillance, London, pp. 21–26, September 5–7, 2007.

[31] Radhakrishnan, R. & Divakaran, A., Systematic acquisition of audio classes for elevator surveillance. *SPIE Image and Video Communications Processing*, **5685**, pp. 64–71, 2005.

[32] Clavel, C., Ehrette T., & Richard, G., Event detection for an audio-based surveillance system. Proceedings of IEEE International Conference on Multimedia and Expo, Amsterdam, pp. 1306–1309, July 2005.

[33] Rouas, J.-L., Louradour, J., & Ambellouis, S., Audio events detection in public transport vehicles. Proceedings of IEEE Intelligent Transportation System Conference, Toronto, pp. 733–738, 2006.

[34] Vacher, M., Istrate, D., Besacier, L., Serignat J.-F., & Castelli, E., Sound detection and classification for medical telesurvey. Proceedings of International Conference of Biomedical Engineering, Innsbruck, pp. 395–399, 2004.

[35] Atrey, P.K., Maddage, N.C., & Kankanhalli, M.S., Audio based event detection for multimedia surveillance. Proceedings of International Conference on Acoustics, Speech and Signal Processing, Toulouse, pp. 813–816, May 2006.

[36] Ito, A., Aiba, A., Ito, M., & Makino, S., Detection of abnormal sound using multistage GMM for surveillance microphone. Proceedings of the Fifth International Conference on Information Assurance and Security, Xian, pp. 733–736, August 18–20, 2009.

[37] Markou, M. & Singh, S., Novelty detection: A review. *Signal Processing, Elsevier*, **83(12)**, pp. 2481–2497, 2003.

[38] Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M., The DET curve in assessment of detection task performance. Proceedings of the Eurospeech, Rhodos, pp. 1895–1898, September 1997.

[39] Bocchetti, G., Flammini, F., Pappalardo, A., & Pragliola, C., Dependable integrated surveillance systems for the physical security of metro railways. Proceedings of the 3rd ACM/IEEE International Conference on Distributed Smart Cameras, Como (Italy), pp. 1–7, August 30 to September 2, 2009.