University of Groningen

Trainable COPE Features for Sound Event Detection

Strisciuglio, Nicola; Petkov, Nicolai

Published in:

Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 24th Iberoamerican Congress, CIARP 2019, Proceedings

DOI:

[10.1007/978-3-030-33904-3_56](https://doi.org/10.1007/978-3-030-33904-3_56)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:

2019

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Strisciuglio, N., & Petkov, N. (2019). Trainable COPE Features for Sound Event Detection. In I. Nyström, Y. Hernández Heredia, & V. Milián Núñez (Eds.), *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 24th Iberoamerican Congress, CIARP 2019, Proceedings* (pp. 599-609). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 11896). Springer. https://doi.org/10.1007/978-3-030-33904-3_56

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Trainable COPE Features for Sound Event Detection

Nicola Strisciuglio^(✉) and Nicolai Petkov

Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence,
University of Groningen, Groningen, Netherlands
n.strisciuglio@rug.nl

Abstract. Systems for automatic analysis of sounds and detection of events are of great importance as they can be used as substitutes of or complement to video analytic systems. In this paper we describe a flexible system for the detection of audio events based on the use of trainable COPE (Combination of Peaks of Energy) features. The structure of a COPE feature is determined in an automatic configuration process on a single prototype example. Thus, they can be adapted to different kinds of sounds of interest. We configure a set of COPE features in order to account for robustness to variations of the characteristics of sounds within a specific class. The proposed system is flexible as new features (also configured on examples drawn from new classes) can be easily added to the feature set. We performed experiments on the MIVIA road events data set for road surveillance applications and compared the results that we achieved with the ones of other existing methods.

Keywords: Few shot training · Sound event detection · Trainable features

1 Introduction

Automatic detection of sound events of interest gained a great attention in the past years [8], because of its various applicative aspects. In social robotics, for instance, sound signals are used to improve the interaction with the environment and people [9]. Security and surveillance are also of great importance and applications of sound analysis to road and traffic monitoring were explored [10]. In the latter case, sound analysis can complement or substitute video analytic systems. Video analytic systems have to deal with varying illumination conditions (e.g. day and night), changing weather conditions, and the presence of occlusions in the field of view of the camera. In some cases the installation of cameras is not allowed due to privacy issues (e.g. a public toilet) or too expensive (e.g. in the monitoring of long roads). Furthermore, certain abnormal events, such as gun shots, screams or skidding tires are very difficult to detect on video only. In [10] a practical system for surveillance of roads by audio analysis was designed as an alternative to video-based analytic systems.

In this paper, we present a method for audio event detection that is based on trainable feature extractors, called COPE (Combination of Peaks of Energy), recently proposed in [27]. The COPE algorithm is based on the analysis of local maxima in a time-frequency representation of the input audio signal, which have been demonstrated to be robust to additive noise [31]. From an applicative point of view, we focus on the case study of event detection for road surveillance.

A comprehensive survey on audio surveillance was published in [8]. Early approaches to sound event detection were based on the extraction of a set of features to describe important properties of the input audio stream. Temporal and frequency features were combined in hand-engineered sets, e.g. energy, pitch, bandwidth, etc. Feature vectors were then used to train a classifier model, subsequently used in the operating phase for the detection and classification of events of interest in unknown audio streams [6, 11, 14, 23, 30]. Subsequently, more complex architectures or data representations were proposed to strengthen the robustness of the system to high variability of the background noise: multi-stage classifiers [19], a classification rejection module [7], representations of the audio signal based on the bag of features approach [12, 20]. In [25], the performance of bag-of-features methods were measured on live audio streams. The temporal arrangement of instantaneous features was taken into account in [15], and in [22] with a pyramidal approach. In [11], instead the audio stream was represented as a time-frequency distribution of its energy and object-detection techniques were employed to detect the events of interest. All these methods involve a feature engineering step, in which a proper set of features for the problem at hand is defined. Such process requires knowledge of the specific domain of application and is critical for the design of a robust recognition system. Deep learning and Convolutional Networks based approaches are able to learn effective features from training data [1, 24], but require large amount of data. Similarly, AENET [28] is based on the VGG architecture and was designed to optimize the trade-off between required computational resources and performance accuracy of sound event detection. In [13], an audio event detection method based on the MobileNet architecture was proposed that outperformed the detection performance of AENET. In the context of audio event recognition, the DCASE challenge is of great importance and focuses on the benchmark of new methods for polyphonic event detection and localization [18] and specific applications in domestic environments [29]. Co-occurring sound detection is also receiving a growing interest [16].

The COPE features that we propose differ from existing approaches as they are trainable with few data, in that they can be configured to detect any sound pattern of interest by presenting a single example to a configuration algorithm. This allows to determine the important features directly from the training data, also when few examples are available, so reducing the effort required by a feature engineering step. The concept of trainable filters was introduced in previous works with successful application to image processing [2, 3] and also integrated into Convolutional Networks [17]. In this work, we deploy the COPE features in a flexible sound recognition system that allows for the extension of the COPE

feature set, and compare the results with its counterpart with a fixed set of COPE features. The results that we achieved on the MIVIA road events data set demonstrate the effectiveness of the COPE features for sound event detection.

2 Method

The event detection method that we propose is based on COPE feature extractors. Given a prototype sound of interest, a COPE feature can be configured by an automatic process that determines a model of the constellation of energy peaks in a time-frequency representation (we use the Gammatonegram) of such sound. In the testing phase, it is then able to detect the prototype sound event and also modified versions of it due to noise or distortion. We configure a set of COPE features on few training sound examples for each class of interest, in order to account for intra-class variability.

2.1 Gammatonegram

We process the input audio signal $s(t)$ by a Gammatone filterbank, which outputs a spectrogram-like image $S_{gt}(t, f)$ called gammatonegram. In Fig. 1a, we show the gammatonegram of a car crash sound. The gammatone filterbank is a linear model of the response of the cochlea membrane of the human auditory system, which converts the incoming sound pressure waves into neural stimuli on the auditory nerve [21].

In contrast with the spectrogram, in which the energy distribution over frequency is computed by band-pass filters all with the same bandwidth, in the gammatonegram the bandwidth of the band-pass filters is linear with their central frequency. Thus, at low frequency the band-pass filters have a narrower bandwidth than the ones at high frequency. This accounts for different resolution at low and high frequency, similarly to the way the auditory system processes the sound. We refer the reader to [21, 27] for details.

2.2 COPE Features

The COPE algorithm receives as input the energy peaks (i.e. local maxima, in a 8-connected neighborhood) extracted from the Gammatonegram of a sound and responds strongly to constellations of points similar to that included in the COPE model during a configuration process performed on a single prototype sound. In Fig. 1b, we show the constellation of energy peaks corresponding to the gammatonegram of the sound in Fig. 1a.

For the configuration of a COPE feature, one has to choose the size of its support, i.e. the size of the time interval in which to consider the position of the energy peak points. Within the support area we take as reference point the peak with the highest energy (circle dot in Fig. 1b) and describe the remaining points by tuples of three parameters $(\Delta t_i, f_i, e_i)$: Δt_i is the temporal offset of the considered point with respect to the reference point, f_i represents the frequency

Fig. 1. Gammatonegram representation of a (a) car crash. We detect and use the (b) constellation of energy local maxima points in the Gammatonegram for the configuration a COPE feature. In the testing phase the COPE feature is applied with a sliding window approach and provides a response during time, which is high when the constellation of energy local maxima is similar to the one determined in the configuration.

channel and e_i is the energy contained in its bandwidth. The set of tuples results in the model $S = \{(\Delta t_i, f_i, e_i) \mid i = 1, \dots, N\}$, where N is the total number of considered points.

In the application phase, given an audio stream of a certain duration, we compute a response for each energy peak included in the model. Such response indicates a score of the presence of each point (i.e. a certain amount of energy in a given frequency band) along the signal to analyse, and is defined as:

$$s_i(t) = \max_{t', f'} \{ \psi(t, f_i) G_{\sigma'}(t', f') \} \quad (1)$$

$$t - \Delta t \leq t' \leq t + \Delta t, f_i - \Delta f \leq f' \leq f_i + \Delta f$$

where $G_{\sigma'}(\cdot, \cdot)$ is a Gaussian weighting function that accounts for some tolerance in the position of the i -th energy peak within the constellation. The tolerance introduced in the application phase makes the filter robust to deformation of the pattern of interest due to distortion or noise. The function $\psi(t, f)$ represents a measure of similarity between the detected energy peak in the unknown audio stream and the one contained in the model of the COPE feature. In this work we consider $\psi(t, f) = S_{gt}(t, f)$, that takes into account only the position of each point in the constellation and its energy content.

We formally define the response of a COPE feature as the geometric mean of the responses of its tuples (i.e. the responses in Eq. 1):

$$r(t) = \left| \left(\prod_{i=1}^{|S|} s_i(t) \right)^{1/|S|} \right|_{th}, \quad (2)$$

where th is a threshold fraction of the highest possible COPE feature response. We determine the value of th for a class of events of interest by cross-validation experiments. We apply the COPE features with a sliding window approach and achieve a continuous response during time. As an example, in Fig. 1c, we show the response of the COPE feature configured in Fig. 1d and applied on the same samples used for configuration.

2.3 Event Detection and Classification

Given M sounds of interest of classes in the set $\{C_0, C_1, \dots, C_{N-1}\}$, we configure a set of M COPE features, one for each such samples. We perform the detection of the events of interest in a time window $W_{[t_1, t_2]}$ that forward shifts on the input audio signal by one third of its length, where t_1 and t_2 are the boundaries of the time interval. In this work, we deploy two different classification approaches. The first approach considers the COPE feature that achieves the highest response in the interval $W_{[t_1, t_2]}$:

$$\hat{r}_i = \max_{t \in W_{[t_1, t_2]}} r_i(t), i = 1, \dots, M \quad (3)$$

and assign to the sound under analysis the class $C_{out} = c(\hat{r}_i)$ that is associated with the COPE feature with highest response over a certain threshold. We determine a threshold value for each class of interest by cross-validation experiments. We refer at this classification approach as *COPE+max*.

The second classification approach involves the use of a multi-class Support Vector Machine (SVM) classifier, trained using the COPE feature vectors constructed by computing the response of the M configured COPE features. We train one SVM for each of the N classes of interest. For the i -th SVM we use the feature vectors computed on the samples of class C_i as positive and the remaining samples as negative. In the classification phase, each SVM provides a score s_i , which we use to determine the class of the sound of interest as:

$$C_{out} = \begin{cases} C_0, & \text{if } s_i < 0 \quad \forall i = 0, \dots, N-1 \\ \arg \max_i s_i, & \text{else.} \end{cases} \quad (4)$$

where C_0 is the background sound class. We refer at this classification approach as *COPE+SVM*.

3 Experiments

3.1 MIVIA Road Events Data Set

We carried out experiments on the MIVIA road events data set [10], which is publicly available for research purpose at the url <http://mivia.unisa.it>. The data set is composed of 57 audio clips of about one minute, sampled at 32 KHz and with a resolution of 16 bits per sample. In order to simulate the occurrence of events of interest in various environments, each audio clip contains a number of events that are superimposed to different combinations of typical background sounds of roads (moving cars, traffic jam, crowds, etc.). The data set contains a total of 400 events of interest, namely 200 car crash and 200 tire skidding events, which occur in different situations with different background sounds. The audio clips are divided in four independent folds each containing 50 events of interest per class. Such organization of the data is suitable for cross-validation experiments. In the following we refer to car crash with *CC*, to tire skidding with *TS* and to background noise with *BN*.

3.2 Performance Evaluation

For the evaluation we adopt the experimental protocol proposed in [12]. We perform the detection of events of interest within a time window of 3 s that forward shifts on the audio stream by steps of 1 s. An event of interest is correctly detected if it is detected in at least one of the time windows that overlap with it. When only background noise is present in the considered time window and an event of interest is wrongly detected, we count it as a false positive. In the case a false detection happens in two consecutive time windows, only one false positive event is counted.

We compute the following measures: the true positive rate (TPR) or recognition rate, error rate (ER), miss detection rate (MDR) and false positive rate (FPR). The MDR and FPR provide information about the robustness of the method to various types of background noise.

Table 1. Classification matrices achieved by the (a) COPE+max and (b) COPE+SVM approaches on the MIVIA road events data set.

<i>COPE+max</i>					<i>COPE+SVM</i>				
		Guessed class					Guessed class		
		TS	CC	Miss			TS	CC	Miss
True	TS	84%	6%	10%	True	TS	92%	2%	6%
class	CC	7%	89.5%	3.5%	class	CC	0.5%	96%	3.5%
(a)					(b)				

3.3 Results and Comparison

In Table 1, we report the classification matrices achieved by *COPE+max* and *COPE+SVM* methods on the MIVIA road events data set. The *COPE+max* approach achieved an average recognition rate of 86.75% with standard deviation 3.4 on the 4-folds cross-validation experiments, while the *COPE+SVM* approach achieved an average recognition rate of 94% with standard deviation 4.3. The rate of missed detection is low (*COPE+max*: 6.75%, *COPE+SVM*: 4.75%) and, coupled with a false positive rate of $FPR = 7.01\%$ for *COPE+max* and 3.94% for *COPE+SVM*, indicates that the proposed method based on trainable COPE features is robust to various types of background noise.

In Table 2, we report the results achieved by the proposed method and compare them with the ones achieved by other existing approaches. We consider the feature sets used in [5, 10] that were combined with classification architectures based on the bag of features (BoF) classification approach. The performance of three sets of features to describe short-time properties of the audio signal were studied and compared. The recognition rate achieved by COPE features is generally higher than the ones achieved by other approaches. Furthermore, the rate of miss detected events is considerably lower (more than 12%) with respect to existing methods. Most of the classification errors are due to misclassification of the target class of the event of interest. The error rate is, indeed, higher than for other methods. From a practical point of view, it means that the occurrence of an abnormal event is however detected and an alert to the human operator could be raised anyway. The FPR achieved by the proposed approach (7.01%) is higher than the method proposed in [5] based on a combination of temporal and spectral, but it is lower than the ones achieved by the system that uses MFCC and Bark feature descriptors for instantaneous properties of the audio signal.

4 Discussion

The system presented in this work, based on a bank of COPE features, achieved high performance results on the MIVIA road events data set. The higher detection capabilities achieved by COPE features with respect to other approaches

Table 2. Comparison of the results achieved on the MIVIA road events data set with respect to existing methods. TPR refers to true positive rate, MDR to miss detection rate, ER to error rate and FPR to false positive rate.

Results comparison				
	TPR	MDR	ER	FPR
COPE + max	86.75%	6.75%	6.5%	7.01%
σ	3.91	4.2	5.23	1.43
COPE + SVM	94%	4.75%	1.25%	3.94%
σ	4.3	4.92	1.25	1.82
BoF [5]	82%	17.75%	0.25%	2.85%
σ	7.79	8.06	1	2.52
BoF (Bark) [10]	80.25%	21.75%	3.25%	10.96%
σ	7.75	8.96	2.5	8.43
BoF (Mfcc) [10]	80.25%	19%	0.75%	7.69%
σ	11.64	11.63	0.96	5.92

are attributable to the fact that they implicitly encode information about the temporal evolution of instantaneous characteristics of the audio stream. This is in contrast with bag of features approaches, where the temporal information is completely lost. The trainable character of the COPE features contributes to the construction of a flexible system for detection of events of interest. Indeed, we configured a number of COPE features on training events of interest (i.e. various car crash and tire skidding events) and used them to detect the occurrence of similar events in unknown audio streams. In principle, one can configure filters that are able to detect other kinds of event and add them to the filterbank in order to extend the detection capabilities of the system. It is worth noting that a COPE feature is configured by only one training sample from which it is able to extract important features, which are later detected in unknown audio streams. This is in contrast with deep learning approaches, which usually require a large number of training samples to learn a robust set of features [4]. In the proposed *COPE+max* system, one can add more COPE feature detectors to an operating detection system without retraining the classifier. The performance of the *COPE+max* are, however, lower than the *COPE+SVM* approach that combines COPE feature vectors with an SVM classifier. In the latter case, the classification system cannot be extended with more feature detectors (e.g. for other classes of sounds of interest) as it need to be retrained. One can, thus, choose a more flexible system in cases where errors can be better handled, while the *COPE+SVM* system is suitable for situations where more precision and reliability is required.

The input features to a COPE feature extractor, namely the energy local maxima in the Gammatonegram of the sound, are reliable and robust to noise characteristic of audio signals [31]. The use of a Gammatone filterbank is not strictly coupled with COPE features, since one can decide to use a time-frequency

representation of the sound that is more suitable for the particular application at hand (i.e. a spectrogram, wavelets, etc.). Moreover, the tolerance introduced in the application phase of a COPE feature accounts for detection of the prototype sound used for configuration and also modified versions of it due to noise and sound distortion. This implies that the proposed filters are robust to noise and locally provide good generalization capabilities. They can be employed together with machine learning techniques to further improve the performance and reliability of the system. In this work, indeed, we considered the filters in the filterbank to give equal contribution in the detection task and determined threshold values by analyzing the ROC curves achieved on the training set. However, different filters could contribute in different ways (i.e. with different weight) to the detection task. Thus, the use of machine learning techniques to determine a set of weights for the filters allows to construct a more robust and effective filterbank. As an example, the filter selection procedure proposed in [26] can be employed to reduce the number of concerned filters in the filterbank by selecting only the most relevant ones. Such procedure provides an optimized filterbank with high effectiveness and lower computational resource demand.

The computation of the response of a COPE feature can be paralleled, since the responses of its tuples are independent from each other. Moreover, given the Gammatonegram transformation of the input audio signal, the responses of the filters in the filterbank can also be implemented on a parallel architecture.

5 Conclusions

In this paper, we proposed a system for audio event detection that is based on COPE features to extract important features from the input audio signals. COPE features are trainable and rely on the detection of specific constellation of energy peaks in a time-frequency representation of the audio signal. Their structure is determined during an automatic configuration process, given prototype audio events of interest. The flexibility introduced by COPE features results in an adaptive system that can be easily extended to detect various other events of interest. We also demonstrated that COPE feature extractors can be used to design more reliable sound recognition systems (with less scalability) by combining their responses with a multi-class classifier. The performance results that we achieved on the MIVIA road events data set are promising and higher than other existing methods. The good generalization capabilities observed in the experiments make the system suitable for use in real-world applications.

References

1. Aytar, Y., Vondrick, C., Torralba, A.: SoundNet: learning sound representations from unlabeled video. In: NIPS 2016 (2016)
2. Azzopardi, G., Petkov, N.: Trainable COSFIRE filters for keypoint detection and pattern recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(2), 490–503 (2013). <https://doi.org/10.1109/TPAMI.2012.106>

3. Azzopardi, G., Strisciuglio, N., Vento, M., Petkov, N.: Trainable COSFIRE filters for vessel delineation with application to retinal images. *Med. Image Anal.* **19**(1), 46–57 (2015)
4. Bengio, Y.: Learning deep architectures for AI. *Found. Trends® Mach. Learn.* **2**(1), 1–127 (2009). <https://doi.org/10.1561/2200000006>
5. Carletti, V., Foggia, P., Percannella, G., Saggese, A., Strisciuglio, N., Vento, M.: Audio surveillance using a bag of aural words classifier. In: *IEEE AVSS*, pp. 81–86 (2013)
6. Clavel, C., Ehrette, T., Richard, G.: Events detection for an audio-based surveillance system. In: *ICME*, pp. 1306–1309 (2005)
7. Conte, D., Foggia, P., Percannella, G., Saggese, A., Vento, M.: An ensemble of rejecting classifiers for anomaly detection of audio events. In: *IEEE AVSS*, pp. 76–81 (2012)
8. Crocco, M., Cristani, M., Trucco, A., Murino, V.: Audio surveillance: a systematic review. *ACM Comput. Surv.* **48**(4), 52:1–52:46 (2016)
9. Do, H.M., Sheng, W., Liu, M.: Human-assisted sound event recognition for home-service robots. *Robot. Biomim.* **3**(1), 7 (2016). <https://doi.org/10.1186/s40638-016-0042-2>
10. Foggia, P., Petkov, N., Saggese, A., Strisciuglio, N., Vento, M.: Audio surveillance of roads: a system for detecting anomalous sounds. *IEEE Trans. Intell. Transp. Syst.* **17**(1), 279–288 (2016). <https://doi.org/10.1109/TITS.2015.2470216>
11. Foggia, P., Saggese, A., Strisciuglio, N., Vento, M.: Cascade classifiers trained on gammatonegrams for reliably detecting audio events. In: *IEEE AVSS*, pp. 50–55 (2014)
12. Foggia, P., Petkov, N., Saggese, A., Strisciuglio, N., Vento, M.: Reliable detection of audio events in highly noisy environments. *Pattern Recogn. Lett.* **65**, 22–28 (2015). <https://doi.org/10.1016/j.patrec.2015.06.026>
13. Foggia, P., Saggese, A., Strisciuglio, N., Vento, M., Vigilante, V.: Detecting sounds of interest in roads with deep networks. In: Ricci, E., Rota Bulò, S., Snoek, C., Lanz, O., Messelodi, S., Sebe, N. (eds.) *ICIAP 2019*. LNCS, vol. 11752, pp. 583–592. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30645-8_53
14. Gerosa, L., Valenzise, G., Tagliasacchi, M., Antonacci, F., Sarti, A.: Scream and gunshot detection in noisy environments. In: *Proceedings of the EURASIP European Signal Processing Conference*, Poznan, Poland (2007)
15. Grzeszick, R., Plinge, A., Fink, G.A.: Temporal acoustic words for online acoustic event detection. In: Gall, J., Gehler, P., Leibe, B. (eds.) *GCPR 2015*. LNCS, vol. 9358, pp. 142–153. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24947-6_12
16. Imoto, K., Kyochi, S.: Sound event detection using graph Laplacian regularization based on event co-occurrence. *CoRR abs/1902.00816* (2019). <http://arxiv.org/abs/1902.00816>
17. López-Antequera, M., Leyva Vallina, M., Strisciuglio, N., Petkov, N.: Place and object recognition by CNN-based COSFIRE filters. *IEEE Access* **7**, 66157–66166 (2019)
18. Mesaros, A., Heittola, T., Virtanen, T.: Metrics for polyphonic sound event detection. *Appl. Sci.* **6**(6), 162 (2016). <https://doi.org/10.3390/app6060162>
19. Ntalampiras, S., Potamitis, I., Fakotakis, N.: An adaptive framework for acoustic monitoring of potential hazards. *EURASIP J. Audio Speech Music Process.* **2009**, 13:1–13:15 (2009). <https://doi.org/10.1155/2009/594103>
20. Pancoast, S., Akbacak, M.: Bag-of-audio-words approach for multimedia event classification. In: *Proceedings of the Interspeech 2012 Conference* (2012)

21. Patterson, R.D., Robinson, K., Holdsworth, J., Mckeown, D., Zhang, C., Allerhand, M.: Complex Sounds and auditory images. In: *Auditory Physiology and Perception*, pp. 429–443 (1992)
22. Plinge, A., Grzeszick, R., Fink, G.A.: A bag-of-features approach to acoustic event detection. In: *IEEE ICASSP*, pp. 3704–3708 (2014)
23. Rabaoui, A., Davy, M., Rossignol, S., Ellouze, N.: Using one-class SVMs and wavelets for audio surveillance. *IEEE Trans. Inf. Forensics Secur.* **3**(4), 763–775 (2008). <https://doi.org/10.1109/TIFS.2008.2008216>
24. Ravanelli, M., Bengio, Y.: Speaker recognition from raw waveform with SincNet. In: *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1021–1028 (2018)
25. Saggese, A., Strisciuglio, N., Vento, M., Petkov, N.: Time-frequency analysis for audio event detection in real scenarios. In: *AVSS*, pp. 438–443, August 2016. <https://doi.org/10.1109/AVSS.2016.7738082>
26. Strisciuglio, N., Azzopardi, G., Vento, M., Petkov, N.: Supervised vessel delineation in retinal fundus images with the automatic selection of B-COSFIRE filters. *Mach. Vis. Appl.* 1–13 (2016). <https://doi.org/10.1007/s00138-016-0781-7>
27. Strisciuglio, N., Vento, M., Petkov, N.: Learning representations of sound using trainable COPE feature extractors. *Pattern Recogn.* **92**, 25–36 (2019). <https://doi.org/10.1016/j.patcog.2019.03.016>
28. Takahashi, N., Gygli, M., Gool, L.V.: AENet: learning deep audio features for video analysis. *IEEE Trans. Multimed.* **20**(3), 513–524 (2018). <https://doi.org/10.1109/TMM.2017.2751969>
29. Turpault, N., Serizel, R., Parag Shah, A., Salamon, J.: Sound event detection in domestic environments with weakly labeled data and soundscape synthesis, June 2019
30. Valenzise, G., Gerosa, L., Tagliasacchi, M., Antonacci, F., Sarti, A.: Scream and gunshot detection and localization for audio-surveillance systems. In: *IEEE AVSS*, pp. 21–26 (2007)
31. Wang, A.L., Th Floor Block F: An industrial-strength audio search algorithm. In: *Proceedings of the 4th International Conference on Music Information Retrieval* (2003)