University of Groningen

Detecting Sounds of Interest in Roads with Deep Networks

Foggia, Pasquale; Saggese, Alessia; Strisciuglio, Nicola; Vento, Mario; Vigilante, Vincenzo

Published in:

Image Analysis and Processing – ICIAP 2019 - 20th International Conference, Proceedings

DOI:

[10.1007/978-3-030-30645-8_53](https://doi.org/10.1007/978-3-030-30645-8_53)

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):

Foggia, P., Saggese, A., Strisciuglio, N., Vento, M., & Vigilante, V. (2019). Detecting Sounds of Interest in Roads with Deep Networks. In E. Ricci, N. Sebe, S. Rota Bulò, C. Snoek, O. Lanz, & S. Messelodi (Eds.), *Image Analysis and Processing – ICIAP 2019 - 20th International Conference, Proceedings* (pp. 583-592). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 11752). Springer Verlag. https://doi.org/10.1007/978-3-030-30645-8_53

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.

Detecting Sounds of Interest in Roads with Deep Networks

Pasquale Foggia¹, Alessia Saggese¹(✉), Nicola Strisciuglio², Mario Vento¹,
and Vincenzo Vigilante¹

¹ Department of Information Engineering,
Electrical Engineering and Applied Mathematics (DIEM),
University of Salerno, Fisciano, Italy

{pfoggia,asaggese,mvento,vvigilante}@unisa.it

² Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence,
University of Groningen, Groningen, The Netherlands
n.strisciuglio@rug.nl

Abstract. Monitoring of public and private places is of great importance for security of people and is usually done by means of surveillance cameras. In this paper we propose an approach for monitoring of roads, to detect car crashes and tire skidding, based on the analysis of sound signals, which can complement or, in some cases, substitute video analytic systems. The system that we propose employs a MobileNet deep architecture, designed to efficiently run on embedded appliances and be deployed on distributed systems for road monitoring. We designed a recognition system based on analysis of audio frames and tested it on the publicly available MIVIA road events data set. The performance results that we achieved (recognition rate higher than 99%) are higher than existing methods, demonstrating that the proposed approach can be deployed on embedded devices in a distributed surveillance system.

Keywords: Audio event detection · Deep learning · MobileNet

1 Introduction

The recent needs for monitoring of cities and public environments stimulated the artificial intelligence and pattern recognition research communities to develop innovative approaches for behaviour analysis [1,6], action recognition [15,34], traffic monitoring [36], among others.

Many existing approaches are based on analysis of visual information recorded by means of surveillance cameras and are successfully employed in monitoring of private and public spaces. The use of cameras has, however, limitations that are determined by occlusions, due to objects in the scene or to the camera field of view, and varying illumination or weather conditions. For instance, algorithms have to deal with changes between daylight and night, or with fog, snow, rain, and so on. In this context, the analysis of audio information has a

complementary role to video analysis, in that the two data modalities can be employed together to increase the reliability of the recognition and surveillance systems. In some cases, such as public toilets for privacy issues, audio analytic systems can be deployed as substitute of video analytic systems. In very large deposit areas or along big roads, the use of microphones instead of cameras can have positive implications on the deployment costs of a safety monitoring system. Microphones are cheaper than cameras and can cover a larger area [13]. In this work, we design a system for detection of events of interest for monitoring of safety of road environments based on efficient deep networks, which can be deployed on embedded systems.

Existing methods for audio surveillance were analyzed in a comprehensive survey paper in [11]. Early works focused on designing sets of features that could extract important information from the input audio signal, such as log-frequency filter banks [26], Mel-frequency Cepstral Coefficients [20], perceptual linear prediction coefficients [30]. In order to design a meaningful feature set, substantial knowledge about the problem at hand and the characteristics of the sounds of interest was necessary. For instance, a peculiar feature of a gunshot is its abrupt increase of signal energy at high frequencies, which is recognizable as a high value of the zero crossing rate feature. Hand-engineered feature sets were used together with a classifier to perform the detection and recognition tasks. Various classifiers were proposed. In [2,9], Gaussian Mixture Model (GMM) based classifiers were deployed, while One-Class Support Vector Machines were proposed in [31]. Spectro-temporal features were also used together with various classifiers, as they are able to effectively describe frequency information in signal sub-bands during time that can be used to distinguish abnormal sounds from common background sounds [8,12]. A biologically-inspired time-frequency representation, called Gammatonegram, was modeled in [29] from measurements performed on the cochlea membrane of the auditory system and subsequently used in [14] for detection of sound of interest with an AdaBoost classifier.

More recently, research was focused on learning suitable representations directly from training data, as engineering robust feature sets requires a broad, not always available, knowledge of the places and sounds of interest, and is not scalable to dynamic environments. Methods for learning representations based on the bag-of-words approach were proposed [3,7,28]. Systems tailored for surveillance applications were subsequently proposed [17], with variants based on non-negative matrix factorization [19] or sparse coding [23]. Applications of these methods to road environments were studied in [13,16] and to changing scenarios in [33]. Biologically-inspired feature extractors, named COPE (Combination of Peaks of Energy), were proposed and are able to learn robust sound representation from single prototype samples [38,39]. Convolutional Networks (CNNs) and Deep learning approaches were more recently proposed to learn effective, hierarchical representations of the sounds of interest jointly with a classification model [4]. The representation power of CNNs and deep networks is a key factor for their success and use in sound event detection [10,21,24].

The recognition method that we propose in this work is based on the convolutional architecture of MobileNet [22,35], thought to be efficient enough to run on embedded and mobile systems, trained to detect events of interest for road monitoring in time-frequency representations of the input audio signal. We report the results of experiments that we performed on the MIVIA road events data set [13], and compare them with those reported by other existing methods. We show that the proposed system achieves good performance, maintaining computational requirements that are feasible for the deployment on embedded devices in a surveillance and monitoring system. Furthermore, differently from other existing deep learning based approaches, is robust enough even if the amount of data for training is not so high.

The rest of the paper is organized as follows. In Sect. 2, we describe the architecture of the proposed method. In Sect. 3.1 we present the data set used for the experiments, while in Sect. 3.2 we report the results that we achieved and compare them with those of other existing approaches. Finally, we draw conclusions in Sect. 4.

2 The Proposed Method

In order to detect sounds with very short duration (like the ones that may arise on the road), we segment the audio stream in consecutive *audio frames* by means of a 3 s mobile Hamming windows. Furthermore, in order to also have the possibility to detect those events that may occur on the border of the windows, a 66% overlap between two consecutive windows has been introduced.

Given a single audio-frame, we employ a time-frequency based representation, which has been widely employed by the scientific community, due to the high representational power; indeed, it has been experimentally shown that this kind of representation generally performs better than still temporal ones (namely the raw data, encoding temporal information of the signal) [21,25]. Among the time-frequency available representations, we decided to use the most famous one, namely the spectrogram. This is a visual representation of the spectrum of frequencies of the sound based on Short-Time Fourier Transform (STFT). An example of the spectrogram for our classes of interest is reported in Fig. 1. Thanks to this choice, we can manage an audio analysis problem as it would be in practice an image analysis problem.

It implies to have the possibility to inherit the wide literature in this field related to deep learning, and in particular to Convolutional Neural Networks (CNNs), which achieved a great success in almost all the image classification tasks. Although so powerful, deep networks have not been widely used in sound event recognition problems, although achieving a notable success in other audio-based tasks, such as music classification [18], sentiment analysis [41] or sentence classification [27].

In more details, our network inherits the architecture of the well known MobileNet network [22,35], introduced so as to be particularly suited for working in real time over embedded systems. Indeed, in a recent benchmarking reporting

Fig. 1. Example of the spectrogram for Car crash (a) and Tire skidding (b) events of interest. X-axis represents the time (in seconds, from 0 to 3 s), while the Y-axis the frequency (in kHz, and it ranges from 0 to 16 with a linear scale).

Fig. 2. Architecture of the MobileNet v2 network used in the proposed system.

the performance of several well-known deep networks applied to the problem of image recognition [5], MobileNet has proved to be a good tradeoff between the accuracy, the number of floating-point operations and the number of parameters required for training the network.

The main idea behind MobileNet lies in the fact that a convolutional layer can be split in a *depthwise* operation followed by a *pointwise* operation, without paying in terms of representational power of the layer. The main advantage before this choice is that 3×3 convolutions require 8 to 9 times less operations, with a consequent reduction in the number of parameters [22]. Having a lower number of parameter in turns also allows to reduce the number of data required for training. This is a very important and not negligible feature when dealing with sound event recognition problem, and in particular with audio analysis on the road, since the amount of data available for training is very limited.

In more details, we use MobileNet v2 [35], whose architecture is reported in Fig. 2. It is equipped with a novel layer, namely the inverted residual with linear bottleneck; this layer takes as input a low-dimensional compressed representa-

Table 1. Details of the composition of the MIVIA road events data set. The duration is expressed in seconds. BN refers to background noise, CC to car crash and TS to tire skidding.

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

tion, which is expanded and then filtered with a lightweight depthwise convolution layer. In other words, the number of feature maps is increased (expansion) and then decreased (projection); data is scattered in a higher-dimensional space so that the non-linear power of ReLU activation function can be exploited without information loss.

3 Experiments

In this section we describe the dataset used in our experimentation (in Subsect. 3.1), together with the results achieved on this dataset, and compare them with the ones achieved by state of the art methodologies (in Subsect. 3.2).

3.1 Dataset

We carried out experiments on the MIVIA road events data set [13], 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 each, sampled at 32 KHz and with a resolution of 16 bits per sample. According to our knowledge, this is the only dataset available in the literature for the problem of sound event recognition on the road. The data set contains 400 events of interest (200 car crash and 200 tire skidding), which occur superimposed to different background sounds. The audio clips are divided in four independent folds each containing 50 events of interest per class, 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*. In Table 1, we report details about the composition of the data set.

3.2 Results and Comparison

The experimentation has been performed by a 4-folds cross validation: in order to reduce the variability on the results, three folds have been used for training and the remaining for testing. Finally, the results obtained over the four tests have been averaged and reported.

In more details, we compute the following metric:

- Recognition Rate (RR), computed as the ratio between the number of correctly classified events of interest and the total number of events.

- False Positive Rate (FPR), computed as the ratio between the number of events detected when only the background is present (false positives) and the total number of events.
- Miss Rate (MR), computed as the ratio between the number of events of interest not detected (misses) and the total number of events.
- Error Rate (ER): computed as the ratio between the number of events of interest detected but misclassified (classification errors) and the total number of events.

Furthermore, we use two different protocols, namely *frame-based* and *event-based*. According to the *frame-based* protocol, the recognition rate, the miss rate and the error rate are computed by considering the total number of audio frames, namely the total number of spectrogram images. Viceversa, in the *event-based* protocol the evaluation is not done by considering the single audio frame but the whole event. In particular, an event of interest (TS and CC), is considered correctly detected if it is identified for at least one of the consecutive windows in which it appears. Performance calculated in this second way will offer a better understanding of the perceived performance of the final system, since it accounts for the fact that, from the point of view of the user, it is sufficient to detect each event once to trigger an alarm, discarding a certain amount of irrelevant False Negatives; this protocol is widely used in literature since it makes up for groundtruth ambiguities regarding the precise start and stop time of an event.

The training of the network has been performed by a mini-batch gradient descent optimization algorithm based on the adaptive learning rate RMSprop [32]. Furthermore, a fine tuning has been performed: ImageNet dataset has been used for training the network and initializing the weights; then, a fine tuning is performed by using Mivia Audio Road dataset.

The results are reported in Table 2; the proposed solution achieve a 88.85% of recognition rate with the frame based protocol, while an impressive 99.50% of recognition rate with the event based protocol.

It is still more impressive that this is not paid in terms of false positives; indeed, we have only a 3.76% of FPR. The results are still more impressive if compared with state of the art methodologies, based both on traditional and deep learning methodologies, as shown in Table 2.

The paper proposed in [17,39] and [13], whose performance ranges from 80.25% to 94.00%, are based on traditional machine learning approaches. The improvement in this case (from about +5% to about +19%) is notable considering the reduced size of the dataset we used for the experimentation. This result, indeed, confirms that MobileNet v2 we used in our system is able to achieve good results even in presence of a limited amount of training data.

This is not the same for other deep learning based architecture, and in particular NASNet [42] and AENET [40]. According to a recent benchmarking [5], NASNet is able to reach the highest accuracy on the problem of image recognition, paying this in terms of time required for the elaboration, being extremely slow. Viceversa, AENet is based on VGG [37], which is according to the same benchmark reported above among the best tradeoff between accuracy and com-

putational burden. Furthermore, AENet is, at the best of our knowledge, the only network proposed in the literature for sound event recognition whose network weights are freely available for benchmarking purposes. The proposed approach outperforms both AENet and NASNet, with an improvement of about +4,5% and +5% in terms of recognition rate over frame based protocol. Such results confirm the effectiveness of the proposed approach.

Table 2. Comparison of the proposed approach with other state of the art methodologies in terms of recognition rate (RR), false positive rate (FPR), miss rate (MR) and error rate (ER) applying the frame by frame and the event based protocols. For each one, we also report whether the Neural Network was trained from scratch or was fine-tuned from a pretrained network in a transfer learning fashion.

Method	Ref.	Training	Frame-based (%)				Event-based (%)			
			RR	FPR	MR	ER	RR	FPR	MR	ER
Proposed	-	fine tuning	88.85	3.76	6.95	4.20	99.50	3.76	0.00	0.50
COPE	[39]	from scratch	—	—	—	—	94.00	3.95	4.75	1.25
HF+BoW	[13]	from scratch	—	—	—	—	82.00	2.85	17.75	0.25
BARK+BoW	[17]	from scratch	—	—	—	—	80.25	7.69	19.00	0.75
MFCC+BoW	[17]	from scratch	—	—	—	—	80.25	10.96	21.75	3.25
NASNet	[42]	fine tuning	84.58	5.51	8.23	7.19	—	—	—	—
AENet	[40]	fine tuning	85.23	6.12	10.14	4.63	—	—	—	—

4 Conclusions

We designed a system for road monitoring and detection of hazardous events by automatic analysis of sound signals. The proposed approach is based on processing a time-frequency representation of the input sounds by means of a lightweight convolutional architecture named MobileNet; as we demonstrated, it is able to achieve high recognition results in highly noisy road environments. The low amount of trainable parameters of MobileNet makes this architecture particularly suitable for problems of road monitoring, where labeled training examples are available in limited quantity. The proposed system achieved higher performance results (event recognition rate > 99%) than other existing methods on the MIVIA road events data set and qualifies to be deployed on embedded distributed surveillance systems, as it has low computational requirements.

References

1. Acampora, G., Foggia, P., Saggese, A., Vento, M.: Combining neural networks and fuzzy systems for human behavior understanding. In: 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance, pp. 88–93, September 2012. <https://doi.org/10.1109/AVSS.2012.25>

2. Atrey, P.K., Maddage, N.C., Kankanhalli, M.S.: Audio based event detection for multimedia surveillance. In: IEEE ICASSP, vol. 5 (2006)
3. Aucouturier, J.J., Defreville, B., Pachet, F.: The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music. *J. Acoust. Soc. Am.* **122**(2), 881–891 (2007)
4. Aytar, Y., Vondrick, C., Torralba, A.: SoundNet: learning sound representations from unlabeled video. In: NIPS 2016 (2016)
5. Bianco, S., Cadene, R., Celona, L., Napoletano, P.: Benchmark analysis of representative deep neural network architectures. *IEEE Access* **1** (2018). <https://doi.org/10.1109/ACCESS.2018.2877890>
6. Brun, L., Saggese, A., Vento, M.: Dynamic scene understanding for behavior analysis based on string kernels. *IEEE Trans. Circ. Syst. Video Technol.* **24**(10), 1669–1681 (2014). <https://doi.org/10.1109/TCSVT.2014.2302521>
7. 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). <https://doi.org/10.1109/AVSS.2013.6636620>
8. Chu, S., Narayanan, S., Kuo, C.C.J.: Environmental sound recognition with time-frequency audio features. *IEEE Trans. Audio Speech Lang. Process.* **17**(6), 1142–1158 (2009). <https://doi.org/10.1109/TASL.2009.2017438>
9. Clavel, C., Ehrette, T., Richard, G.: Events detection for an audio-based surveillance system. In: ICME, pp. 1306–1309 (2005). <https://doi.org/10.1109/ICME.2005.1521669>
10. Colangelo, F., Battisti, F., Carli, M., Neri, A., Calabró, F.: Enhancing audio surveillance with hierarchical recurrent neural networks. In: AVSS, pp. 1–6, August 2017. <https://doi.org/10.1109/AVSS.2017.8078496>
11. Crocco, M., Cristani, M., Trucco, A., Murino, V.: Audio surveillance: a systematic review. *ACM Comput. Surv.* **48**(4), 52:1–52:46 (2016). <https://doi.org/10.1145/2871183>
12. Dennis, J., Tran, H.D., Chng, E.S.: Image feature representation of the subband power distribution for robust sound event classification. *IEEE Trans. Audio Speech Lang. Process.* **21**(2), 367–377 (2013)
13. Foggia, P., Petkov, N., Saggese, A., Strisciuglio, N., Vento, M.: Audio surveillance of roads: a system for detecting anomalous sounds. *IEEE Intell. Transp. Syst.* **17**(1), 279–288 (2016). <https://doi.org/10.1109/TITS.2015.2470216>
14. Foggia, P., Saggese, A., Strisciuglio, N., Vento, M.: Cascade classifiers trained on gammatonegrams for reliably detecting audio events. In: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 50–55, August 2014
15. Foggia, P., Saggese, A., Strisciuglio, N., Vento, M.: Exploiting the deep learning paradigm for recognizing human actions. In: IEEE AVSS 2014, pp. 93–98 (2014). <https://doi.org/10.1109/AVSS.2014.6918650>
16. Foggia, P., Saggese, A., Strisciuglio, N., Vento, M., Petkov, N.: Car crashes detection by audio analysis in crowded roads. In: IEEE AVSS, pp. 1–6 (2015). <https://doi.org/10.1109/AVSS.2015.7301731>
17. 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>
18. Fu, Z., Lu, G., Ting, K.M., Zhang, D.: A survey of audio-based music classification and annotation. *IEEE Trans. Multimed.* **13**(2), 303–319 (2011). <https://doi.org/10.1109/TMM.2010.2098858>

19. Giannoulis, D., Stowell, D., Benetos, E., Rossignol, M., Lagrange, M., Plumbley, M.D.: A database and challenge for acoustic scene classification and event detection. In: EUSIPCO, pp. 1–5 (2013)
20. Guo, G., Li, S.Z.: Content-based audio classification and retrieval by support vector machines. *IEEE Trans. Neural Netw.* **14**(1), 209–215 (2003)
21. Hertel, L., Phan, H., Mertins, A.: Comparing time and frequency domain for audio event recognition using deep learning. In: International Joint Conference on Neural Networks (IJCNN), pp. 3407–3411 (2016)
22. Howard, A.G., et al.: MobileNets: efficient convolutional neural networks for mobile vision applications. *arXiv* (2017)
23. Lu, X., Tsao, Y., Matsuda, S., Hori, C.: Sparse representation based on a bag of spectral exemplars for acoustic event detection. In: IEEE ICASSP, pp. 6255–6259 (2014). <https://doi.org/10.1109/ICASSP.2014.6854807>
24. Medhat, F., Chesmore, D., Robinson, J.: Environmental sound recognition using masked conditional neural networks. In: Cong, G., Peng, W.-C., Zhang, W.E., Li, C., Sun, A. (eds.) ADMA 2017. LNCS (LNAI), vol. 10604, pp. 373–385. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69179-4_26
25. Mesaros, A., et al.: Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge. *IEEE/ACM Trans. Audio Speech Lang. Process.* **26**(2), 379–393 (2018). <https://doi.org/10.1109/TASLP.2017.2778423>
26. Nadeu, C., Macho, D., Hernandez, J.: Time and frequency filtering of filter-bank energies for robust HMM speech recognition. *Speech Commun.* **34**, 93–114 (2001). [https://doi.org/10.1016/S0167-6393\(00\)00048-0](https://doi.org/10.1016/S0167-6393(00)00048-0)
27. Ouyang, X., Gu, K., Zhou, P.: Spatial pyramid pooling mechanism in 3D convolutional network for sentence-level classification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **26**(11), 2167–2179 (2018). <https://doi.org/10.1109/TASLP.2018.2852502>
28. Pancoast, S., Akbacak, M.: Bag-of-audio-words approach for multimedia event classification. In: Interspeech (2012)
29. 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)
30. Portelo, J., Bugalho, M., Trancoso, I., Neto, J., Abad, A., Serralheiro, A.: Non-speech audio event detection. In: IEEE ICASSP, pp. 1973–1976 (2009)
31. Rabauoi, 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)
32. Ruder, S.: An overview of gradient descent optimization algorithms. *CoRR* abs/1609.04747 (2016). <http://arxiv.org/abs/1609.04747>
33. Saggese, A., Strisciuglio, N., Vento, M., Petkov, N.: Time-frequency analysis for audio event detection in real scenarios. In: IEEE AVSS, pp. 438–443 (2016). <https://doi.org/10.1109/AVSS.2016.7738082>
34. Saggese, A., Strisciuglio, N., Vento, M., Petkov, N.: Learning skeleton representations for human action recognition. *Pattern Recogn. Lett.* **118**, 23–31 (2019). <https://doi.org/10.1016/j.patrec.2018.03.005>
35. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *arXiv* (2018)
36. Shirazi, M.S., Morris, B.T.: Looking at intersections: a survey of intersection monitoring, behavior and safety analysis of recent studies. *IEEE Trans. Intell. Transp. Syst.* **18**(1), 4–24 (2017). <https://doi.org/10.1109/TITS.2016.2568920>

37. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)
38. Strisciuglio, N., Vento, M., Petkov, N.: Bio-inspired filters for audio analysis. In: Amunts, K., Grandinetti, L., Lippert, T., Petkov, N. (eds.) BrainComp 2015. LNCS, vol. 10087, pp. 101–115. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50862-7_8
39. 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>
40. 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>
41. Zhang, S., Zhang, S., Huang, T., Gao, W., Tian, Q.: Learning affective features with a hybrid deep model for audio-visual emotion recognition. IEEE Trans. Circ. Syst. Video Technol. **28**(10), 3030–3043 (2018). <https://doi.org/10.1109/TCSVT.2017.2719043>
42. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. CoRR abs/1707.07012 (2017). <http://arxiv.org/abs/1707.07012>