Temporal Acoustic Words for Online Acoustic Event Detection

Rene Grzeszick, Axel Plinge, and Gernot A. Fink
{*rene.grzeszick, axel.plinge, gernot.fink*}*@tu-dortmund.de*

Department of Computer Science, TU Dortmund

Abstract. The Bag-of-Features principle proved successful in many pattern recognition tasks ranging from document analysis and image classification to gesture recognition and even forensic applications. Lately these methods emerged in the field of acoustic event detection and showed very promising results. The detection and classification of acoustic events is an important task for many practical applications like video understanding, surveillance or speech enhancement. In this paper a novel approach for online acoustic event detection is presented that builds on top of the Bag-of-Features principle. Features are calculated for all frames in a given window. Applying the concept of feature augmentation additional temporal information is encoded in each feature vector. These feature vectors are then softly quantized so that a Bag-of-Feature representation is computed. These representations are evaluated by a classifier in a sliding window approach. The experiments on a challenging indoor dataset of acoustic events will show that the proposed method yields state-of-the-art results compared to other online event detection methods. Furthermore, it will be shown that the temporal feature augmentation significantly improves the recognition rates.

1 Introduction

The detection and classification of acoustic events is an important task for many practical applications. In analysis of multimedia content, the classification of objects, visual actions or movements and sounds can be combined for the understanding high level semantic events in videos [9]. It is also possible to do this multimedia event classification based on acoustic features alone [13]. Live applications include the analysis of acoustic events in various environments. Surveillance in cluttered scenes can be improved by an acoustic analysis in order to detect unexpected scenarios that are not visually recognizable (e.g. screams or glass breaking) [2, 5]. Another application is meeting analysis and multi-modal interaction [20]. A slightly different field are outdoor applications like mobile robots for security, urban planning [26, 21] or wildlife observations where the goal is to determine the presence of certain animals by acoustic features [10, 27]. The task is difficult because of the diversity of the acoustic events. A single event is usually comprised of a variety of individual sounds, e.g. chair movement can produce knocking and rubbing sounds, handling paper can include rustling and

knocking on the table and so on. Human laughter or speech are fundamentally different depending on the individual person. It is desirable for a classification method of acoustic events to handle these variabilities and generalize from single instances to the broad range of sounds within an event class.

In order to capture the temporal variability of different sounds, HMMs are widely used. However, the Viterbi decoding requires a full sequence in order to predict the past [4]. Consequentially, most HMM approaches work offline and assign event classes to time points for a past sequence of events. Thus they commonly only address the task of offline analysis.

There are several methods for online classification and detection of acoustic events. The basic method is to use a GMM to model each category, as is done in speaker identification. The mean and variance of the feature vector are modelled as Gaussians. This is also known as the *Bag-of-Frames* approach to acoustic classification [1, 6]. Extensions of this approach include the use of a background model [24]. Lately, methods that build on the Bag-of-Features principle have emerged in the field of acoustic event detection [2, 13, 16]. Acoustic features such as MFCCs are extracted for each frame and clustered in order to build a set of representatives. The occurrences of these representatives in a short time window are then counted and the resulting histogram is used for classification. An very similar approach is the so called superframe, where a histogram over a pre-classification is used instead [15, 14]. Given the task at hand, these representatives are often referred to as an audio or acoustic word. One advantage of the Bag-of-Features models is that due to their simplicity and fast computation it is easy to employ them for online analysis.

The basic Bag-of-Features approach employs unsupervised hard vector quantization in order to derive a codebook by which to quantize the input [13]. This strategy is not always optimal for acoustic classification. It is rather advantageous to follow the GMM approach of using soft quantization by assuming a Gaussian distribution of the feature vectors and perform the training in a supervised manner [16], which is termed *Bag-of-Super-Features*.

These approaches discard any temporal information within the analysis window by treating all frames with disregard of temporal order. One way to reintroduce temporal information is to use a pyramid scheme [11]. The short time windows that are used for classification are well suited for a subdivision as proposed by the pyramid scheme [16]. In contrast to the pyramid scheme there are approaches in computer vision that propose directly including this information at feature level [8, 17]. This is sometimes referred to as feature augmentation.

In [17] features are augmented with continuous x, y coordinates that encode the position of a feature within an image. This directly builds on the encoding abilities of the Fisher Vector approach. Given a set of features, a GMM is estimated in order to compute a set of representatives, e.g. visual or, here, acoustic words. These represent the global distribution of the samples. For the encoding, each feature is assigned to the visual/acoustic words based on the GMM posteriors. Then, the differences of the local distribution with respect the global distribution of the acoustic words are encoded the mean and covariance devia-

tion vectors of the feature vector and the visual/acoustic word. While this allows to append continuous coordinates and yields a very detailed encoding compared to a hard or soft quantization, it also requires enough samples in order to robustly estimate the local distributions. Given the low number of frames in a time window this is hardly possible in acoustic classification and event detection.

In [8] quantized x, y coordinates that roughly encode the position of a feature within an image are appended. This approach preserves a tiling structure similar to the pyramid scheme and does not estimate the local feature distributions. The augmentation of the features with quantized coordinates causes the clustering step in the Bag-of-Feature computation to form different codebooks for different regions of an image or a time window. It could be shown that these adaptive codebooks cover the information contained in each tile better than a global codebook and allow for reducing the dimensionality of the representation.

In this paper it will be shown that the detection and classification of acoustic events based on Bag-of-Super-Features representations of acoustic words can be improved by augmenting the features with a temporal component. The evaluation will show that a tiling with adaptive codebooks as proposed in [8] outperforms plain Bag-of-Features methods as well as pyramid schemes in recognition rates while at the same time having a lower dimensionality. Furthermore, the evaluation will show the influence of parameters such as window length and codebook size on the Bag-of-Super-Features approach and finally a comparison with recent methods will show that the proposed approach achieves state-of-the-art results.

2 Method

For the acoustic event detection and classification, a single microphone or beamformed signal is processed in short time windows of w seconds. An overview of the processing method is shown in Fig. 2. For a given window i, a set of feature vectors $Y_i = (\boldsymbol{y}_1 \dots \boldsymbol{y}_K)$ is calculated for all K frames in this window. All features in this set are augmented with additional temporal information with respect to the window. These features are then softly quantized by a GMM that has been trained in a supervised manner so that a Bag-of-Features representation is computed. Finally, a multinomial maximum likelihood classifier is applied.

2.1 Features

For sound and especially speech processing, the mel frequency cepstral coefficients (MFCCs) are one of the most widely used features. The input signal is filtered by a triangular mel frequency filter bank. In the computational modeling of the human hearing process [25], ERB-spaced gammatone filterbanks are used. From that the gammatone frequency cepstral coefficients (GFCCs) were

A video of the proposed method applied in our lab can be found at: https://vimeo.com/134489154

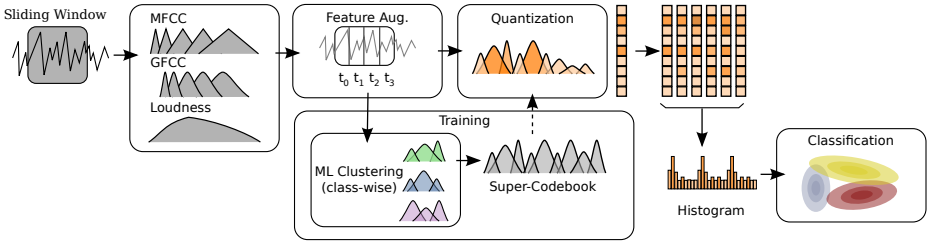

Fig. 1. Overview of the method: Given a window containing an acoustic signal, MFCCs, GFCCs and a loudness feature are computed. The resulting feature vector is augmented by a quantized time coordinate with respect to the window. A GMM is applied for clustering the features of each class in order to learn a supervised codebook. Finally, all features are quantized and the resulting histogram is classified by a multinomial maximum likelihood classifier.

derived [19]. The filterbank of the MFCCs is replaced by linear phase gammatone filters. The basic feature vector is comprised of regular MFCCs, GFCCs and the perceptual loudness derived from the A-weighted magnitude spectrum. A basic whitening step is performed by subtracting the mean and dividing by the standard deviation of the traning data.

2.2 Feature Augmentation

It has been shown that adding time information is able to improve the recognition rates of acoustic event detection and classification [16]. This idea is highly related to the encoding of spatial information in the vision domain [11]. In contrast to the popular pyramid approach, in the following, the time information is directly encoded at feature level [8].

Therefore, quantized time coordinates t are appended to the feature vector. Given a fixed window of w seconds in length, it is subdivided into N tiles of equal size so that the time is quantized into a value of $[1, .., N]$. Thus the augmented feature vector consists of 13 MFCCs m, 13 GFCCs g, loudness l and a temporal index t:

$$\boldsymbol{y}_k = (m_1, \ldots, m_{13}, g_1, \ldots, g_{13}, l, t)^T \tag{1}$$

Note that when quantizing these features by a vector quantizer or a GMM in order to compute a Bag-of-Features representation this generates adaptive codebooks for each tile. This is a major difference to the spatial pyramid approach where the same codebook is used for each tile. Furthermore, the size of the codebook V determines the size of the overall feature representation whereas the size of the feature representation grows with each tile in the pyramid scheme [8, 11]. In this approach only tiling is used and the upper levels of the pyramid are discarded as they usually do not carry much information (cf. [8]).

published in: German Conference on Pattern Recognition (GCPR) 2015

2.3 Bag-of-Super-Features

After augmenting the feature vectors with temporal information, a Bag-of-Features approach is applied. Hence, a codebook of *acoustic words* is estimated from the training set. Most Bag-of-Features approaches use clustering algorithms, e.g. k-means, on the complete training set to derive a codebook and later assign each feature to a centroid by hard quantization.

However, disregarding the labels in the clustering step can lead to mitigation of significant differences. A remedy for this effect is to build codebooks of size Z for all C classes Ω_c separately and then to concatenate them into a large super-codebook. This method is referred to as a *Bag-of-Super-Features* (cf. [16]) in analogy to the super-vector construct used in speaker identification [22].

Here, the expectation-maximization (EM) algorithm is applied to all feature vectors $\boldsymbol{y}_k$ for each class Ω_c in order to estimate Z means and deviations $\mu_{z,c}, \sigma_{z,c}$ for all C classes. All means and deviations are concatenated into a super-codebook $\boldsymbol{v}$ with $V = Z \cdot C$ elements

$$v_{j=(c \cdot Z+z)} = (\mu_{z,c}, \sigma_{z,c}) \tag{2}$$

where the index j is computed from the class index c and the Gaussian index z as $j = c \cdot Z + z$. Using this super-codebook, a soft quantization of a feature vector $\boldsymbol{y}_k$ can be computed as

$$q_{k,j}(\boldsymbol{y}_k, v_j) = \mathcal{N}(\boldsymbol{y}_k | v_j) \; / \; \sum_{j'} \mathcal{N}(\boldsymbol{y}_k | v_{j'}) \; . \tag{3}$$

Then, a histogram $\boldsymbol{b}$ can be computed over all K frames of an input window Y_i, where the occurrences of an acoustic word v_j in the window Y_i are estimated by

$$b_i(Y_i, v_j) = \frac{1}{K} \sum_k q_{k,j}(\boldsymbol{y}_k, v_j) \; . \tag{4}$$

These histograms can then be used as a feature representation of the window Y_i and as an input for a classifier.

2.4 Classification

The probability of an acoustic word v_j to occur in a given class Ω_c is estimated using a set of training windows $Y_i \in \Omega_c$ for each class c by Laplacian smoothing:

$$P(v_j | \Omega_c) = \frac{\alpha + \sum_{Y_i \in \Omega_c} b_i(Y_i, v_j)}{\alpha V + \sum_{u=1}^{V} \sum_{Y_m \in \Omega_c} b_m(Y_m, v_u)} \; , \tag{5}$$

where α is weighting factor for the smoothing (in practice $\alpha = 0.5$ showed good results). Hence, the probability is estimated by the fraction of the acoustic word v_j to occur in any window of class c with respect to all acoustic words occurring in any window class c. Rather than using a prior classification step to eliminate

silence and background noise, as done in several systems (cf. [23]), the rejection class Ω_0 is trained with recordings where no event occurred.

Since all classes are assumed to be equally likely and have the same prior, maximum likelihood classification is used. The posterior is estimated using the relative frequency of all acoustic words

$$P(Y_i|\Omega_c) = \prod_{v_j \in \boldsymbol{v}} P(v_j|\Omega_c)^{b_i(Y_i,v_j)} \ . \tag{6}$$

For the classification of a single window Y_i the maximum probability is chosen for deriving a label that is assigned to this window.

2.5 Detection

Due to the simplicity and rapid computation of this approach it can easily be adapted to event detection. Here, a sequence of acoustic events is given.

The classification window is applied as a sliding window that is moved forward for one frame k at a time. The recognition result is used for the frame that is centered in the window so that context information is available for a short time before and after the frame. As the window has a length of w seconds, there is a processing delay of $w/2$ seconds. As the implementation is running in real time, this delay is of high interest. In the experiments it will be shown that a delay of 300 ms is sufficient for practical purposes.

3 Evaluation

The proposed method has been evaluated on the very challenging office live task of the DCase (Detection and Classification of Acoustic Scenes and Events) challenge [6]. The temporal feature augmentation is compared with a Bag-of-Super-Features approach without feature augmentation and the pyramid scheme. Parameters with respect to temporal processing, like the windows size and tilings, as well as the influence of the codebook size are evaluated. The approach is then compared to the state of the art methods. In order to test for significant differences between classifiers and parameter configurations, a randomization test ($N = 1e5$) has been performed [7]. This method was chosen since it does avoid any distribution assumption.

3.1 D-Case office live dataset

The dataset of this task is comprised of a variety of indoor sounds that could occur in an office or comparably a meeting room scenario. There are 16 sound classes *alert, clearthroat, cough, doorslam, drawer, keyboard, knock, laughter, mouse, pageturn, pendrop, phone, printer, speech, switch, keys* and additionally *silence* that have to be detected. The dataset provides a training set of segmented sequences for each of the 16 classes with a total length of 18 minutes and 49

		tilings					
	w	2	4	6	8	10	12
temporal	0.3	50.6±3.9	50.8±3.5	50.8±3.4	50.6±3.3	50.7±3.4	50.6±3.4
	0.6	53.8±3.0	55.3±2.6	55.7±3.1	55.7±3.0	55.4±2.8	55.5±2.7
	0.9	51.7±5.9	53.5±4.2	55.2±4.5	55.3±4.6	55.1±3.9	55.2±4.2
	1.2	50.0±5.3	52.0±3.7	53.1±4.6	54.2±4.4	54.1±3.9	54.1±3.6
	1.5	43.1±9.1	48.3±6.3	49.9±6.4	51.1±6.2	51.8±5.8	52.0±5.1
pyramid	0.3	50.3±3.6	50.1±3.5	50.0±3.5	49.7±3.5	49.5±3.4	49.6±3.4
	0.6	54.9±3.1	54.7±2.8	54.6±2.8	54.4±2.7	54.4±2.8	53.9±2.9
	0.9	54.6±3.9	54.2±3.8	54.0±3.9	53.8±4.0	53.6±4.1	53.5±4.2
	1.2	54.3±3.6	54.3±3.5	53.9±3.4	53.9±3.3	53.6±3.3	53.4±3.3
	1.5	50.7±5.4	50.6±5.1	50.2±5.0	50.1±5.0	49.7±5.0	49.4±5.0

Table 1. F-scores [%] and standard deviation for pyramids and temporal feature augmentation for different window lengths and tilings. The results are averaged over all three scripts, both annotations and 50 codebook generations using $Z = 30$.

seconds. Furthermore, there are three scripted test sequences which are publicly available with a total length of 5 minutes and 21 seconds. For each of these sequences two annotations are available. Since there is no training data for the silence/background class, the silence portions from the other two scripts were used to train the classifier for each script. The task is to detect the acoustic events in these sequences and classify them correctly. Hence, for different methods the precision and recall with respect to the number of frames that are correctly recognized are computed and the F-score is evaluated. All experiments were repeated 50 times using different codebooks each time over all sequences and annotations, yielding a total of 300 runs. Note that the differences in the scripts lead to a larger variance as the results for each script differ by about 3%.

3.2 Temporal processing

For the detection of acoustic events, two parameters are of interest with respect to the temporal processing. The first one is the length of the window w in seconds, the second one is the spatial setup within this window, i.e. the number of tiles. The F-scores of different window lengths and tilings for the temporal feature augmentation and the temporal pyramid scheme, as proposed in [16], are shown in Tab. 1. For the pyramids an additional max pooling step has been computed on top of the tilings. All parameter combinations have been evaluated using $Z = 30$, i.e. a super-codebook size of $V = 30 \cdot 17$.

It can be seen that for both methods, the best results are achieved by using a window length of 0.6 s. Furthermore, a baseline method with no spatial information has been evaluated with different window lengths as well. Again the best classification performance of 55.0±3.1% has been achieved with a window length of 0.6 s. The results also show that the adaptive codebooks that are computed for each tile by the feature augmentation approach allow for a more fine grained

Classifier \ Z	20	30	40	60	90	120
feature augmentation	54.3±2.8	55.7±3.1	55.9±3.5	54.8±4.4	51.5±4.9	48.0±4.8
pyramid	54.3±3.2	54.9±3.1	55.1±3.2	54.5±4.0	52.2±4.6	48.9±4.7
w/o temp processing	54.4±2.7	55.0±3.1	54.8±3.3	54.1±4.1	51.6±5.0	48.0±4.9

Table 2. F-scores [%] and standard deviation for pyramids and temporal feature augmentation for different codebook sizes. Best performing temporal configurations are used. Results are averaged over all three scripts, both annotations and 50 codebooks.

analysis. The best results are achieved by using 6 or 8 tiles, while the pyramid scheme shows the best result with only two tiles.

Using the best configuration for each augmentation scheme, the permutation test has been performed. This revealed that the temporal augmentation significantly ($p < 0.01$) outperformed the pyramid and the unaugmented classification. It also showed that the pyramid did not outperform the unaugmented version.

3.3 Codebook size

Different codebook sizes of $Z = 20, 30, 40, 60, 90, 120$ were evaluated for the pyramid approach, the temporal feature augmentation and a Bag-of-Super-Features approach without temporal information. For all methods the best performing temporal processing configurations are used. Hence, a window size of 0.6 s is used for all three approaches. For the pyramid two tiles and for the acoustic words with temporal feature augmentation six tiles are computed.

While for the augmented features the size of the overall feature representation is equal to the super-codebook size $V = C \cdot Z$, the concatenation in the pyramid scheme further increases the size of the final representation. Hence, a temporal pyramid with N tiles at the bottom and one top layer has a final feature representation of the size $(N + 1) \cdot V$.

In Tab. 2 the results are shown. It can be observed that small codebooks of 30 or 40 acoustic words per class yields good results and that the performance deteriorates with an increasing codebook size. The best performance is achieved using temporal feature augmentation and a codebook size that uses 40 centroids per class (i.e. a super-codebook size of $V = 680$).

3.4 Comparison with state-of-the-art

For comparison, some state-of-the-art methods were re-implemented and used in combination with the MFCC-GFCC features. Additionally, published results for the D-Case office live development set were used for comparing the performance.

Re-implemented methods The Bag-of-Frames method [1], the Bag-of-Audio words method [13] and a Bag-of-Features approach using Fisher encoding and a linear SVM (cf. [3]) were evaluated. The Bag-of-Frames estimates one GMM

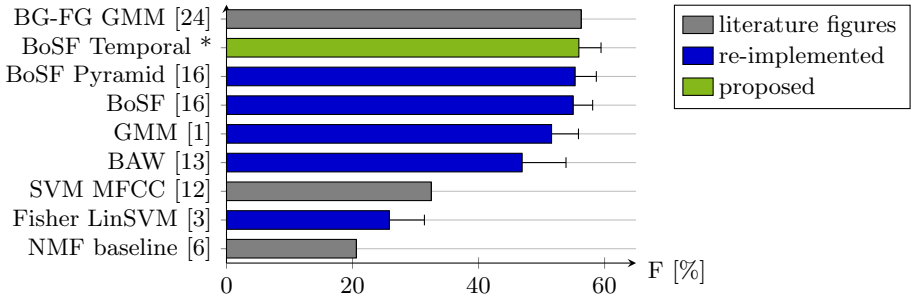

Fig. 2. Comparison of different classifiers and literature values on the D-Case office live development set with the proposed approach * as F-scores [%]. The re-implemented results are averaged over all three scripts, both annotations and 50 codebook generations. The best parameter configuration for each classifier was chosen.

per class. It achieves the best performance with a codebook size of $Z = 30$ per GMM. The Bag-of-Audio words uses hard vector quantization with $V = 1000$ as originally proposed and an SVM with a histogram intersection kernel. As Fisher encoding usually uses smaller codebooks, the best performance was achieved with a codebook size of $Z = 5$ and encoding the mean and covariance deviation vectors. Detailed results are shown in Fig. 2 in blue. The Bag-of-Audio words achieved an F-score of only 47% , which is most likely due to the unsupervised codebook learning. Also the Fisher approach yields an F-score of only 26%. This clearly demonstrates that short time windows do not cover enough frames in order to robustly estimate the local distributions around each centroid of the codebook. With an F-Score of 56% the temporal augmentation outperformed the well known Bag-of-Audio-Words method. The difference was proven significant ($p < 0.01$) by the permutation test.

Results from the literature When comparing these results with the ones published for the D-Case office live development set, shown in Fig. 2 in gray, it can be seen that the temporal augmentation outperforms most live detection methods. Note however, that it is difficult to accurately compare to these results as the protocol might deviate with respect to the number of runs or even more importantly scripts or annotations used in the evaluation. The offline HMM based results are not shown since the task of online detection is investigated. Typically, the best performing offline HMM approaches achieve a 20% higher F-score (cf. [18]). The best performing online method is the GMM based approach using a separate background model [24]. With an F-score of 56.3% it is well in the range of our proposed method. However, the authors state that it is not robust to noise.

3.5 Result discussion

Figure 3 shows the class-wise F-Score over all sequences. The most difficult categories include switch and mouse, which usually last only a few ms and are

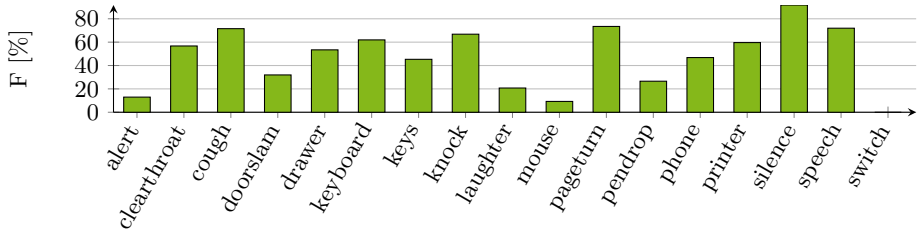

Fig. 3. Classwise F-Score on D-Case Develpment set sequences using the proposed method.

therefore very difficult to detect in an online detection approach that relies on some context. In most cases those are mistaken for silence. Longer lasting (e.g. printing) or very distinctive acoustic events (e.g. knocks or coughs) are more easily recognized. An exemplary result on the first 60 s of a sequence is shown in Fig. 4.

4 Conclusion

In this paper a novel method for online acoustic event detection has been proposed. It builds on the Bag-of-Features principle and integrates feature augmentation with a temporal component and a supervised codebook learning step.

The experiments on a challenging indoor dataset of acoustic events show that the proposed method yields state-of-the-art results compared to other online event detection methods. Furthermore, it could be shown that the feature augmentation yields significant improvements over a basic Bag-of-Features approach and the well known pyramid scheme, while at the same time reducing the dimensionality of the representation. The results show that for practical purposes a processing delay of only 300 ms allows for the integration of enough context to robustly recognize acoustic events.

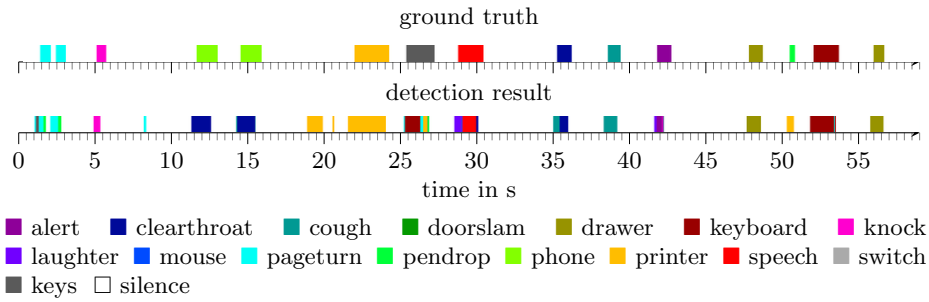

Fig. 4. Example detection results for the first 60 s of sequence 01 of the D-Case office live development set using the proposed method with temporal feature augmentation using six tiles over a window size of 0.6 s and a codebook size of $Z = 40$.

References

1. 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. The Journal of the Acoustical Society of America 122(2), 881–891 (2007)
2. Carletti, V., Foggia, P., Percannella, G., Saggese, A., Strisciuglio, N., Vento, M.: Audio Surveillance using a Bag of Aural Words Classifier. In: 2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance. pp. 81–86. IEEE (Aug 2013)
3. Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: Proc. British Machine Vision Conference (BMVC) (2011)
4. Fink, G.A.: Markov Models for Pattern Recognition, From Theory to Applications. Advances in Computer Vision and Pattern Recognition, Springer, London, 2 edn. (2014)
5. Foggia, P., Saggese, A., Strisciuglio, N., Vento, M.: Cascade classifiers trained on Gammatonegrams for reliably detecting Audio Events. In: Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on. pp. 50–55. IEEE (2014)
6. Giannoulis, D., Benetos, E., Stowell, D., Rossignol, M., Lagrange, M., Plumbley, M.D.: Detection and Classification of Acoustic Scenes and Events: An IEEE AASP Challenge. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). pp. 1–4. IEEE (2013)
7. Good, P.: Permutation Tests – A Practical Guide to Resampling Methods for Testing Hypotheses. Springer Series in Statistics, Springer, 2 edn. (2000)
8. Grzeszick, R., Rothacker, L., Fink, G.A.: Bag-of-Features Representations using Spatial Visual Vocabularies for Object Classification. In: Proc. Int. Conf. on Image Processing (ICIP) (2013)
9. Jiang, Y.G., Bhattacharya, S., Chang, S.F., Shah, M.: High-level event recognition in unconstrained videos. International Journal of Multimedia Information Retrieval 2(2), 73–101 (2013)
10. Klinck, H., Stelzer, K., Jafarmadar, K., Mellinger, D.K.: AAS Endurance: An Autonomous Acoustic Sailboat for Marine Mammal Research. In: Int. Robotic Sailing Conference. Matosinhos, Portugal (Jul 2009)
11. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). vol. 2, pp. 2169–2178 (2006)
12. Nogueira, W., Roma, G., Herrera, P.: Automatic Event Classification using Front End Single Channel Noise Reduction, MFCC Features and a Support Vector Machine Classifier. Tech. rep., IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events (2013), http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/abstracts/OL/NR2.pdf
13. Pancoast, S., Akbacak, M.: Bag-of-Audio-Words Approach for Multimedia Event Classification. In: Interspeech. pp. 2105–2108 (2012)
14. Phan, H., Maasz, M., Mazur, R., Mertins, A.: Random Regression Forests for Acoustic Event Detection and Classification. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2014), http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=6949625

published in: German Conference on Pattern Recognition (GCPR) 2015

15. Phan, H., Mertins, A.: Exploiting Superframe Cooccurence for Acoustic Event Recognition. In: European Signal Processing Conference (2014)
16. Plinge, A., Grzeszick, R., Fink, G.A.: A Bag-of-Features Approach to Acoustic Event Detection. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (2014)
17. Sánchez, J., Perronnin, F., De Campos, T.: Modeling the spatial layout of images beyond spatial pyramids. Pattern Recognition Letters 33(16), 2216–2223 (2012)
18. Schröder, J., Cauchi, B., Schädler, M.R., Moritz, N., Adiloglu, K., Anemüller, J., Doclo, S., Kollmeier, B., Goetze, S.: Acoustic event detection using signal enhancement and spectro-temporal feature extraction. Tech. rep., IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events (2013), http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/abstracts/OL/SCS.pdf
19. Shao, Y., Srinivasan, S., Wang, D.: Incorporating auditory feature uncertainties in robust speaker identification. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. pp. 277–280 (2007)
20. Shivappa, S.T., Trivedi, M.M., Rao, B.D.: Audiovisual Information Fusion in HumanComputer Interfaces and Intelligent Environments: A Survey. Proceedings of the IEEE 98(10), 1692–1715 (Oct 2010)
21. Steele, D., Krijnders, J.D., Guastavino, C.: The Sensor City Initiative: Cognitive Sensors for Soundscape Transformations. GIS Ostrava (2013)
22. Tang, H., Chu, S.M., Hasegawa-Johnson, M., Huang, T.S.: Partially supervised speaker clustering. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(5), 959–971 (2012)
23. Temko, A., Malkin, R., Zieger, C., Macho, D., Nadeu, C., Omologo, M.: CLEAR Evaluation of Acoustic Event Detection and Classification Systems. In: Stiefelhagen, R., Garofolo, J. (eds.) Multimodal Technologies for Perception of Humans, Lecture Notes in Computer Science, vol. 4122, pp. 311–322. Springer Berlin Heidelberg (2007), http://dx.doi.org/10.1007/978-3-540-69568-4_29
24. Vuegen, L., Broeck, B.V.D., Karsmakers, P., Gemmeke, J.F., Vanrumste, B., Hamme, H.V.: An MFCC-GMM Approach for Event Detection and Classification. Tech. rep., IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events (2013), http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/abstracts/OL/VVK.pdf
25. Wang, D., Brown, G.J. (eds.): Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. IEEE Press (2006)
26. Young, S.H., Scanlon, M.V.: Robotic Vehicle uses Acoustic Array for Detection and Localization in Urban Environments. SPIE Proc. Mobile Robot Perception 4364, 264–273 (Sep 2001)
27. Zeppelzauer, M., Stöger, A.S., Breiteneder, C.: Acoustic detection of elephant presence in noisy environments. In: Proceedings of the 2nd ACM international workshop on Multimedia analysis for ecological data. pp. 3–8. ACM (2013)

published in: German Conference on Pattern Recognition (GCPR) 2015