IDENTIFY, LOCATE AND SEPARATE: AUDIO-VISUAL OBJECT EXTRACTION IN LARGE VIDEO COLLECTIONS USING WEAK SUPERVISION

Sanjeel Parekh, Alexey Ozerov, Slim Essid, Ngoc Duong, Patrick Pérez, Gael Richard

IDENTIFY, LOCATE AND SEPARATE: AUDIO-VISUAL OBJECT EXTRACTION IN LARGE VIDEO COLLECTIONS USING WEAK SUPERVISION

Sanjeel Parekh[⋆] *Alexey Ozerov*[†] *Slim Essid*[⋆] *Ngoc Q. K. Duong*[†] *Patrick Pérez*[§] *Gaël Richard*[⋆]

[⋆] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[†] InterDigital, Cesson Sévigné, France [§] Valeo.ai, Paris, France

ABSTRACT

We tackle the problem of audio-visual scene analysis for weakly-labeled data. To this end, we build upon our previous audio-visual representation learning framework to perform object classification in noisy acoustic environments and integrate audio source enhancement capability. This is made possible by a novel use of non-negative matrix factorization for the audio modality. Our approach is founded on the multiple instance learning paradigm. Its effectiveness is established through experiments over a challenging dataset of music instrument performance videos. We also show encouraging visual object localization results.

Index Terms— Audio-visual event detection, source separation, non-negative matrix factorization, multiple instance learning

1. INTRODUCTION

Extracting information from audio-visual (AV) data about events, objects, and scenes finds important application in several areas such as video surveillance, multimedia indexing and robotics. Among other tasks, automatic analysis of AV scenes entails: (i) identifying events or objects, (ii) localizing them in space and time, and (iii) extracting the audio source of interest from the background. In our efforts to build a unified framework to deal with these challenging problems, we presented a first system tackling event identification and AV localization in an arXiv technical report earlier [1].[1] Continuing to build upon that study, in this paper we focus on making event/object classification robust to noisy acoustic environments and incorporating the ability to enhance or separate the object in the audio modality.

There is a long history of works on supervised event detection [2, 3, 4, 5]. However, scaling supervision to large video collections and obtaining precise annotations for multiple tasks is both time consuming and error prone [6, 7]. Hence, in our previous work [1] we resort to training with weak labels *i.e.* global video-level object labels without any timing information. Multiple instance learning (MIL) is a well-known learning paradigm central to most studies using weak supervision [8]. MIL is typically applied to cases where labels are available over bags (sets of instances) instead of individual instances. The task then amounts to jointly selecting appropriate instances and estimating classifier parameters. For applying this to our case, let us begin by viewing a video as a labeled bag, containing a collection of image regions (also referred to as image proposals) and audio segments (also referred to as audio proposals) obtained by chunking the audio temporally. While such a formulation yields promising results using deep MIL [1, 9], its audio proposal design has two limitations with respect to our goals: it is (i) prone to erroneous classification in noisy acoustic conditions and (ii) limited to temporal localization of the audio event or object, thus does not allow for time-frequency segmentation in order to extract the audio source of interest. To address these, we propose to generate audio proposals using non-negative matrix factorization (NMF) [10]. Note that the term *proposal* refers to image or audio "parts" that may potentially contain the object of interest. For the audio modality these "parts" can be obtained through uniform chunking of the signal, as we did previously, or more sophisticated methods.

NMF is a popular unsupervised audio decomposition method that has been successfully utilized in various source separation systems [11] and as a front-end for audio event detection systems [12, 13]. It factorizes an audio spectrogram into two nonnegative matrices namely, so-called spectral patterns and their activations. Such a part-based decomposition is analogous to breaking up an image into constituent object regions. This motivates its use in our system. It makes it possible not only to de-noise the audio, but also to appropriately combine the parts for separation. An interesting work which has appeared recently uses NMF basis vectors with weak supervision from visual modality to perform audio source separation [14]. There are three key differences with our proposed approach: (i) The authors of that proposal use the NMF basis vectors and not their activations for training the system. Hence no temporal information is utilized. (ii) Unlike us, they perform a supervised dictionary construction step after training to decompose a test signal (iii) Finally, they do not consider the task of visual localization. Other recent approaches for deep learning based vision-guided audio source separation methods utilize ground-truth source masks for training [15, 16]. It is worth noting that our proposed enhancement technique is significantly different as we do not use separated ground truth sources at any stage and only rely on weak labels. This makes the problem considerably more challenging.

Contributions. We show how a deep MIL framework can be flexibly used to robustly perform several AV scene understanding tasks using just weak labels. In particular, in addition to temporal audio proposals as in our earlier study [1], we propose to use NMF components as audio proposals for improved classification and to allow source enhancement. We demonstrate the usefulness of such an approach on a large dataset of unconstrained musical instrument performance videos. As the data is noisy, we expect NMF decomposition to provide additional, possibly "cleaner" information about the source of interest. Moreover, scores assigned to each component by the MIL module to indicate their relevance for classification can be reliably used to enhance or separate multiple sources.

We begin with a discussion of various modules of the proposed approach from proposal generation to classification in Sec. 2. This is followed by qualitative and quantitative results on classification, audio source enhancement and visual localization tasks in Sec. 3.

[1] Technical report [1] has not been published in any official proceedings. It was only presented as an extended abstract at a CVPR 2018 workshop.

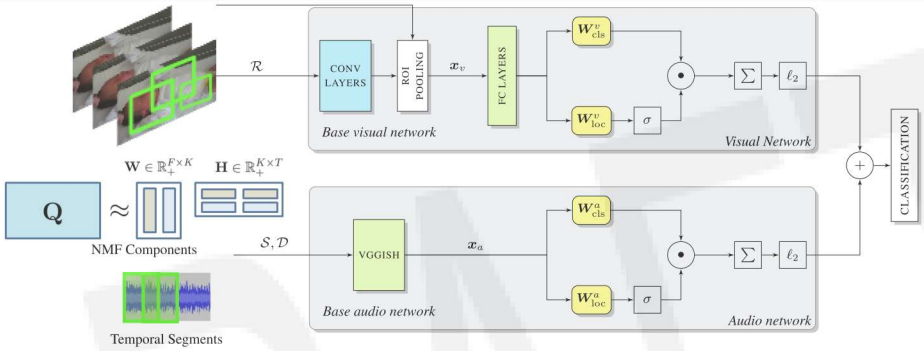

Figure 1: **Proposed approach**: Given a video, we consider the depicted pipeline to go from audio and visual proposals to localization and classification. For the visual modality box proposals are considered, while for audio temporal segments and/or NMF component proposals are utilized. Weights for each module are either trained from scratch (in yellow), fine-tuned (in green) or frozen (in blue) during training.

2. PROPOSED APPROACH

The proposed approach is depicted in Fig. 1. We formulate the problem within a deep MIL framework. Each video is considered as a bag of visual and audio proposals. These proposals are fed to their respective feature extraction and scoring networks. The scores indicate relevance of each region or segment for a particular class. Their aggregation, as depicted in Fig. 1, allows video-level classification. In the following section we discuss proposal generation, feature extraction, scoring and training procedures in detail.

2.1. System Details

Visual Proposals. As our goal is to localize spatially and temporally the most discriminative image region pertaining to a class, we choose to generate proposals over video frames sub-sampled at a rate of 1 frame per second. Class-agnostic bounding-box proposals are obtained using the well-known EdgeBoxes [17] algorithm. To reduce the computational load and redundancy, the confidence score generated by this method is used to select top M_{img} proposals from each sampled image. Hence, for a 10 sec. video, such a procedure would generate a list of $M = 10 \times M_{\text{img}}$ region proposals.

A fixed-length feature vector, $\boldsymbol{x}_v(r_m; V) \in \mathbb{R}^{d_v}$ is obtained from each image region proposal, r_m in a video V, using a convolutional neural network altered with a region-of-interest (RoI) pooling layer. In practice, feature vectors $\boldsymbol{x}_v(\cdot)$ are passed through two fully connected layers, which are fine-tuned during training. Typically, standard CNN architectures pre-trained on ImageNet [18] classification are used for the purpose of initializing network weights (see Sec. 3 for implementation details).

Audio Proposals. We study two kinds of proposals:

1. **Temporal Segment Proposals (TSP)**: Herein the audio is simply decomposed into T temporal segments of equal length, $\mathcal{S} = \{s_1, s_2, \ldots, s_T\}$. These proposals are obtained by transforming the raw audio waveform into log-Mel spectrogram and subsequently chunking it by sliding a fixed-length window along the temporal axis. The dimensions of this window are chosen to be compatible with base audio network (see Sec. 3).
2. **NMF Component Proposals (NCP)**: Using NMF we decompose audio magnitude spectrogram $\mathbf{Q} \in \mathbb{R}_+^{F\times N}$ consisting of F frequency bins and N short-time Fourier transform (STFT) frames, such that,

$$\mathbf{Q} \approx \mathbf{WH}, \tag{1}$$

where $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$ are interpreted as the nonnegative audio spectral patterns and their temporal activation matrices respectively. Here K is the total number of spectral patterns. To estimate $\mathbf{W}$ and $\mathbf{H}$ we minimize the Kullback-Leibler (KL) divergence using multiplicative update rules [10] where $\mathbf{W}$ and $\mathbf{H}$ are initialized randomly.

We now apply NMF-based Wiener filtering, as in [19], to an audio recording to decompose it into K tracks (also referred to as NMF components) each obtained from $\mathbf{W}_k, \mathbf{H}_k$ for $k \in [1, K]$, where $\mathbf{W}_k$ and $\mathbf{H}_k$ denote spectral patterns and activations corresponding to the k^{th} component, respectively. They can now be considered as proposals that may or may not belong to the class of interest. Specifically, we chunk each NMF component into temporal segments, which we call NMF Component proposals or NCPs. We denote the set of NCPs by $\mathcal{D} = \{d_{k,t}\}$, where each element is indexed by the component, $k \in [1, K]$ and temporal segment $t \in [1, T]$ number. As the same audio network is used for both kinds of audio proposals, for each NMF component or track we follow the TSP computation procedure. However, this is done with a non-overlapping window for reducing computational load.

Proposals generated by both the aforementioned methods are passed through a VGG-style deep network known as `vggish` [20] for base audio feature extraction. Hershey *et al.* introduced this state-of-the-art audio feature extractor as an audio counterpart to networks pre-trained on ImageNet for classification. `vggish` has been pre-trained on a preliminary version of YouTube-8M [21] for audio classification based on video tags. It generates a 128 dimensional embedding $\boldsymbol{x}_a(s_t; V) \in \mathbb{R}^{128}$ for each input log-Mel spectrogram segment $s_t \in \mathbb{R}^{96\times 64}$ with 64 Mel-bands and 96 temporal frames. We fine-tune all the layers of `vggish` during training.

Proposal Scoring and Fusion. Having obtained representations for each proposal in both the modalities, we now score them with respect to classes using the two-stream architecture put forth by Bilen *et al.* [22]. This module consists of parallel classification and localization streams. Generically denoting audio or visual proposals by $\mathcal{P}$ and their $l-$dimensional input representations to the

scoring module by $Z \in \mathbb{R}^{|\mathcal{P}|\times l}$, the following sequence of operations is carried out: First, Z is passed through linear fully-connected layers of both classification and localization streams (shown with yellow in Fig. 1) giving transformed matrices $A \in \mathbb{R}^{|\mathcal{P}|\times C}$ and $B \in \mathbb{R}^{|\mathcal{P}|\times C}$, respectively. This is followed by a softmax operation on B in the localization stream, written as:

$$[\sigma(B)]_{pc} = \frac{e^{b_{pc}}}{\sum_{p=1}^{|\mathcal{P}|} e^{b_{pc}}}, \ \forall (p,c) \in (1,|\mathcal{P}|) \times (1,C). \quad (2)$$

This allows the localization layer to choose most relevant proposals for each of the C classes. Subsequently, the classification stream output A is weighted by $\sigma(B)$ through element-wise multiplication: $E = A \odot \sigma(B)$. Class scores over the video are obtained by summing the resulting weighted scores in E over all the proposals.

The same set of operations is carried out for both audio and visual proposals. Before addition of global level scores from both the modalities, they are ℓ_2-normalized to ensure similar score range.

Classification Loss and Training. Given a set of N training videos and labels, $\{(V^{(n)}, \boldsymbol{y}^{(n)})\}_{n=1}^{N}$, we solve a multi-label classification problem. Here $\boldsymbol{y} \in \mathcal{Y} = \{-1,+1\}^C$ with the class presence denoted by $+1$ and absence by -1. To recall, for each video $V^{(n)}$, the network takes as input a set of image regions $\mathcal{R}^{(n)}$ along with audio TSP $\mathcal{S}^{(n)}$, NCP $\mathcal{D}^{(n)}$ or both. After performing the described operations on each modality separately, the ℓ_2 normalized scores are added and represented by $\phi(V^{(n)}; \boldsymbol{w}) \in \mathbb{R}^C$, with all network weights and biases denoted by $\boldsymbol{w}$. Both sub-modules are trained jointly using the multi-label hinge loss:

$$L(\boldsymbol{w}) = \frac{1}{CN} \sum_{n=1}^{N} \sum_{c=1}^{C} \max\Big(0, 1 - y_c^{(n)} \phi_c(V^{(n)}; \boldsymbol{w})\Big). \quad (3)$$

2.2. Source Enhancement

As noted earlier, a by-product of training with NCPs is the ability to perform source enhancement. This can be done by aggregating the NMF component proposal relevance scores as follows:

- Denoting by $\beta_{k,t}$ the score for k^{th} component's t^{th} temporal segment, we compute a global score for each component as

$$\alpha_k = \max_{t \in T} \ \beta_{k,t}.$$

  Note that other temporal aggregation methods such as global average or weighted rank pooling [23] may also be considered. However, our preliminary experiments showed no significant difference in the results while using any of these methods.

- Then we apply min-max scaling between [0,1]. This allows using the aggregated component scores as weights for masking.

$$\alpha'_k = \frac{\alpha_k - \min(\alpha)}{\max(\alpha) - \min(\alpha)}.$$

- This is followed by soft mask based source and noise spectrogram reconstruction using complex-valued mixture spectrogram $\mathbf{X}$. Note that we can optionally apply a hard threshold τ on α'_k to choose the top ranked components for the source. This amounts to replacing α'_k by the indicator function $\mathbf{1}[\alpha'_k \geq \tau]$ in the following reconstruction equations:

$$\mathbf{S} = \frac{\sum_k \alpha'_k \mathbf{W}_k \mathbf{H}_k}{\mathbf{WH}} \mathbf{X}, \quad \mathbf{N} = \frac{\sum_k (1-\alpha'_k) \mathbf{W}_k \mathbf{H}_k}{\mathbf{WH}} \mathbf{X}$$

  Here $\mathbf{S}$ and $\mathbf{N}$ are the estimates of source of interest and of background noise, respectively.

3. EXPERIMENTS

3.1. Setup

Dataset. We use Kinetics-Instruments (KI), a subset of the Kinetics dataset [24] that contains 10s Youtube videos from 15 music instrument classes. From a total of 10,267 videos, we create training and testing sets that contain 9199 and 1023 videos, respectively. For source enhancement evaluation, we handpicked 45 "clean" instrument recordings, 3 per class. Due to their unconstrained nature, the audio recordings are mostly noisy, *i.e.* videos are either shot with accompanying music/instruments or in acoustic environments containing other background events. In that context, "clean" refers to solo instrument samples with minimal amount of such noise.

Systems. Based on the configuration depicted in Fig. 1, we propose to evaluate audio-only, A, and audio-visual (multimodal), V + A, systems with different audio proposal types, namely:

- A (TSP): temporal segment proposals,
- A (NCP): NMF component proposals,
- A (TSP, NCP): all TSPs and NCPs are put together into the same bag and fed to the audio network.

Systems using only TSP already give state-of-the-art results [1], and serve as a strong baseline for establishing the usefulness of NCPs in classification. For source enhancement we compare with the following NMF related methods:

- Supervised NMF [25]: We use the class labels to train separate dictionaries of size 100 for each music instrument with stochastic mini-batch updates. At test time, depending on the label, the mixture is projected onto the appropriate dictionary for source reconstruction.
- NMF Mel-Clustering [26]: This blind audio-only method reconstructs source and noise signals by clustering mel-spectra of NMF components. We take help of the example code provided online for implementation in MATLAB [27].

Implementation Details. All proposed systems are implemented in Tensorflow. They were trained for 10 epochs using Adam optimizer with a learning rate of 10^{-5} and a batch size of 1. We use the MATLAB implementation of EdgeBoxes for generating image region proposals, obtaining approximately 100 regions per video with $M_{img} = 10$. Base visual features $\boldsymbol{x}_v \in \mathbb{R}^{9216}$ are extracted using `caffenet` with pre-trained ImageNet weights and 6×6 RoI pooling layer modification [28]. The fully connected layers, namely fc_6 and fc_7, are fine-tuned with 50% dropout.

For audio, each recording is resampled to 16 kHz before processing. We use the official Tensorflow implementation of `vggish` [29]. The whole audio network is fine-tuned during training. For TSP generation we first compute log-Mel spectrum over the whole file with a window size of 25ms and 10ms hop length. The resulting spectrum is chunked into segment proposals using a 960ms window with a 480ms stride. For log-Mel spectrum computation we use the accompanying `vggish` code implementation. For a 10 second recording, this yields 20 segments of size 96×64. For NCP, we consider $K = 20$ components with KL divergence and multiplicative updates. As stated in Sec. 2.1, each NMF component is passed through the TSP computation pipeline with a non-overlapping window, giving a total of 200 (20×10) NCPs for a 10s audio recording.

Classification at Test Time: Kinetics-Instruments is a multi-class dataset. Hence, we consider $\text{argmax}_c s_c$ of the score vector to be the predicted class and report the overall accuracy

Source Enhancement Evaluation Protocol: We corrupt the original audio with background noise corresponding to recordings of environments such as bus, busy street, park, etc. using one audio file per scene from the DCASE 2013 scene classification dataset [30]. The system can be utilized in two modes: *label known* and *label unknown*. For the former, where the source of interest is known, we simply use the proposal ranking given by the corresponding classifier for reconstruction. For the latter, the system's classification output is used to infer the source.

3.2. Classification Results

In Table 1 we show classification results on KI for all systems explained previously. For methods using NMF decomposition, the accuracy is averaged over 5 runs to account for changes due to random initialization. We observe that the accuracies are consistent across runs *i.e.* the standard deviation does not exceed 0.5 for any of the proposed systems.

	System	Accuracy (%)
(a)	V–only	63.0
(b)	A (TSP)	75.3
(c)	A (NCP)	71.1
(d)	A (NCP, TSP)	76.7
(e)	(b) + (c)	**77.3**
(f)	V + A (TSP)	84.5
(g)	V + A (NCP)	80.9
(h)	V + A (NCP, TSP)	**84.6**
(i)	(f) + (g)	**84.6**

Table 1: Classification results on KI test set. Here, (e) adds the scores of systems (b) and (c) at test time [resp. for (i)]

First, we note an evident increase in performance for all the AV systems when contrasted with audio-only methods. Indeed, the image sequence provides strong complementary information about an instrument's presence when audio is noisy. Also, observe that using NCP in conjunction with TSP results in a noticeable improvement for the audio-only systems. In comparison, this relative difference is negligible for AV methods. A possible explanation is that NCPs are expected to provide complementary information in noisy acoustic conditions. Thus, their contribution in assisting TSP is visible for audio-only classification. On the other hand, vision itself serves as a strong supporting cue for classification, unaffected by noise in audio and its presence limits the reliance on NCP. The accuracy drop when using NCP alone is expected as whole audio segments could often be easier to classify than individual components.

SNR (dB)	V + A (TSP)	V + A (NCP, TSP)
5	78.7	**79.3**
0	73.9	**75.6**
-10	63.2	**65.2**
-20	58.7	**59.2**

Table 2: Classification accuracy on KI dataset for different levels of noise in the test audio

To further test the usefulness of NCP, we corrupt the test set audio with additional noise at different SNRs using samples from DCASE 2013 scene classification data. Average classification scores over this noisy test set are reported in Table 2. We observe a clear improvement even for the AV system when used with NCPs.

Figure 2: Visual localization for different instruments (clockwise from top left: accordion, bagpipes, trombone and saxophone) from the test set. Max. scoring bounding box shown in green.

3.3. Source Enhancement Results and Visual Localization

Following the testing protocol stated in Sec. 3.1, we report, in Table 3, average Source to Distortion Ratio (SDR) [31] over 450 audio mixtures created by mixing each of the 45 clean samples from the dataset with 10 noisy audio scenes. The results look promising but not state-of-the-art. This performance gap can be explained by noting that the audio network is trained for the task of audio event detection and thus does not yield optimal performance for source enhancement. The network focuses on discriminative components, failing to separate some source components from the noise by a larger margin, possibly requiring manual thresholding for best results. Also, performance for the proposed systems does not degrade when used in "Label Unknown" mode, indicating that despite incorrect classification the system is able to cluster acoustically similar sounds. Performance of supervised NMF seems to suffer due to training on a noisy dataset. We present some visual localization examples in Fig. 2. Examples and supplementary material are available on our companion website.[2]

System	Label Known	Label Unknown
Supervised NMF	2.32	–
NMF Mel-Clustering	–	**4.32**
V + A (NCP), soft	3.29	3.29
V + A (NCP), $\tau = 0.1$	**3.77**	3.85
V + A (NCP), $\tau = 0.2$	3.56	3.56
V + A (NCP, TSP), soft	2.11	2.15

Table 3: Average SDR over mixtures created by combining clean instrument examples with environmental scenes.

4. CONCLUSION

We have presented a novel system for robust AV object extraction under weak supervision. Unlike previous multimodal studies, we only use weak labels for training. The central idea is to perform MIL over a set of audio and visual proposals. In particular, we propose the use of NMF for generating audio proposals. Its advantage for robust AV object classification in noisy acoustic conditions and source enhancement capability is demonstrated over a large dataset of musical instrument videos.

[2] http://bit.ly/2IV6XAs

5. REFERENCES

[1] S. Parekh, S. Essid, A. Ozerov, N. Q. K. Duong, P. Pérez, and G. Richard, "Weakly supervised representation learning for unsynchronized audio-visual events," *CoRR*, vol. abs/1804.07345, 2018.

[2] A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen, "Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations," in *ICASSP*. IEEE, 2015, pp. 151–155.

[3] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543–1551, 2010.

[4] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *ICASSP*. IEEE, 2017, pp. 771–775.

[5] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Feature learning with matrix factorization applied to acoustic scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1216–1229, 2017.

[6] J. Xu, A. G. Schwing, and R. Urtasun, "Learning to segment under various forms of weak supervision," in *CVPR*, 2015, pp. 3781–3790.

[7] M. Gao, Z. Xu, L. Lu, A. Wu, I. Nogues, R. M. Summers, and D. J. Mollura, "Segmentation label propagation using deep convolutional neural networks and dense conditional random field," in *ISBI*. IEEE, 2016, pp. 1265–1268.

[8] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[9] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *ECCV*, September 2018.

[10] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[11] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[12] V. Bisot, S. Essid, and G. Richard, "Overlapping sound event detection with supervised nonnegative matrix factorization," in *ICASSP*. IEEE, 2017, pp. 31–35.

[13] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound event detection in multisource environments using source separation," in *Machine Listening in Multisource Environments*, 2011.

[14] R. Gao, R. Feris, and K. Grauman, "Learning to separate object sounds by watching unlabeled video," in *ECCV*, September 2018.

[15] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *ECCV*, September 2018.

[16] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 112:1–112:11, July 2018. [Online]. Available: http://doi.acm.org/10.1145/3197517.3201357

[17] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *ECCV*. Springer, 2014, pp. 391–405.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[19] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[20] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "CNN architectures for large-scale audio classification," in *ICASSP*. IEEE, 2017, pp. 131–135.

[21] S. Abu-El-Haija, N. Kothari, J. Lee, A. P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8M: A large-scale video classification benchmark," in *arXiv:1609.08675*, 2016. [Online]. Available: https://arxiv.org/pdf/1609.08675v1.pdf

[22] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *CVPR*, 2016, pp. 2846–2854.

[23] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 695–711.

[24] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[25] C. Févotte, E. Vincent, and A. Ozerov, "Single-channel audio source separation with NMF: divergences, constraints and algorithms," in *Audio Source Separation*. Springer, 2018, pp. 1–24.

[26] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," in *in Proceedings of International Conference on Digital Audio Effects DAFx09*, 2009.

[27] *NMF Mel Clustering Code*, http://www.ient.rwth-aachen.de/cms/dafx09/.

[28] R. Girshick, "Fast R-CNN," in *ICCV*. IEEE, 2015, pp. 1440–1448.

[29] *Vggish Code*, https://github.com/tensorflow/models/tree/master/research/audioset.

[30] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[31] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.