Apprentissage de représentations pour l'analyse robuste de scènes audiovisuelles

Sanjeel Parekh

université PARIS-SACLAY

TELECOM ParisTech

technicolor

NNT : 2019SACLT015

Thèse de doctorat

Learning representations for robust audio-visual scene analysis

Thèse de doctorat de l'Université Paris-Saclay
préparée à Télécom ParisTech

Ecole doctorale n°580 Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Télécom ParisTech, le 18 mars 2019, par

SANJEEL PAREKH

Composition du Jury :

Laurent Girin Professeur, Grenoble INP, France	Président
Josef Sivic Directeur de Recherche, École Normale Supérieure / Inria, France	Rapporteur
Tuomas Virtanen Professeur, Tampere University of Technology, Finland	Rapporteur
Hervé Bredin Chargé de Recherche, CNRS-LIMSI, France	Examinateur
Nancy Bertin Chargée de Recherche, CNRS - IRISA, France	Examinatrice
Slim Essid Professeur, Télécom ParisTech, France	Directeur de thèse
Gaël Richard Professeur, Télécom ParisTech, France	Co-directeur de thèse
Patrick Pérez Directeur Scientifique, Valeo.ai, France	Co-directeur de thèse
Alexey Ozerov Senior Research Scientist, Technicolor, France	Co-encadrant de thèse
Ngoc Q. K. Duong Senior Research Scientist, Technicolor, France	Co-encadrant de thèse

Abstract

The goal of this thesis is to design algorithms that enable robust detection of objects and events in videos through joint audio-visual analysis. This is motivated by humans' remarkable ability to meaningfully integrate auditory and visual characteristics for perception in noisy scenarios. To this end, we identify two kinds of natural associations between the modalities in recordings made using a single microphone and camera, namely motion-audio correlation and appearance-audio co-occurrence.

For the former, we use audio source separation as the primary application and propose two novel methods within the popular non-negative matrix factorization framework. The central idea is to utilize the temporal correlation between audio and motion for objects/actions where the sound-producing motion is visible. The first proposed method focuses on soft coupling between audio and motion representations capturing temporal variations, while the second is based on cross-modal regression. We segregate several challenging audio mixtures of string instruments into their constituent sources using these approaches.

To identify and extract many commonly encountered objects, we leverage appearance–audio co-occurrence in large datasets. This complementary association mechanism is particularly useful for objects where motion-based correlations are not visible or available. The problem is dealt with in a weakly-supervised setting wherein we design a representation learning framework for robust AV event classification, visual object localization, audio event detection and source separation.

We extensively test the proposed ideas on publicly available datasets. The experiments demonstrate several intuitive multimodal phenomena that humans utilize on a regular basis for robust scene understanding.

Résumé

L'objectif de cette thèse est de concevoir des algorithmes qui permettent la détection robuste d'objets et d'événements dans des vidéos en s'appuyant sur une analyse conjointe de données audio et visuelle. Ceci est inspiré par la capacité remarquable des humains à intégrer les caractéristiques auditives et visuelles pour améliorer leur compréhension de scénarios bruités. À cette fin, nous nous appuyons sur deux types d'associations naturelles entre les modalités d'enregistrements audiovisuels (réalisés à l'aide d'un seul microphone et d'une seule caméra), à savoir la corrélation mouvement/audio et la co-occurrence apparence/audio.

Dans le premier cas, nous utilisons la séparation de sources audio comme application principale et proposons deux nouvelles méthodes dans le cadre classique de la factorisation par matrices non négatives (NMF). L'idée centrale est d'utiliser la corrélation temporelle entre l'audio et le mouvement pour les objets / actions où le mouvement produisant le son est visible. La première méthode proposée met l'accent sur le couplage flexible entre les représentations audio et de mouvement capturant les variations temporelles, tandis que la seconde repose sur la régression intermodale. Nous avons séparé plusieurs mélanges complexes d'instruments à cordes en leurs sources constituantes en utilisant ces approches.

Pour identifier et extraire de nombreux objets couramment rencontrés, nous exploitons la co-occurrence apparence/audio dans de grands ensembles de données. Ce mécanisme d'association complémentaire est particulièrement utile pour les objets où les corrélations basées sur le mouvement ne sont ni visibles ni disponibles. Le problème est traité dans un contexte faiblement supervisé dans lequel nous proposons un framework d'apprentissage de représentation pour la classification robuste des

événements audiovisuels, la localisation des objets visuels, la détection des événements audio et la séparation de sources.

Nous avons testé de manière approfondie les idées proposées sur des ensembles de données publics. Ces expériences permettent de faire un lien avec des phénomènes intuitifs et multimodaux que les humains utilisent dans leur processus de compréhension de scènes audiovisuelles.

Abstract's french translation revised by Martin Engilberge.

Acknowledgements

Many things must fall into place to be able to delve into something so vast and unknown and still emerge smiling after three years, ready to experience it once more. Indeed, it took a great set of people and places to make this work possible.

I would first like to thank my advisors: Alexey, Gaël, Ngoc, Patrick and Slim. It would not have been possible without their guidance, encouragement and support. I am extremely grateful to them for giving me the freedom to choose and walk on research paths of my choice. To me, the five of them were multimodality personified!

I am really thankful to Josef Sivic and Tuomas Virtanen for agreeing to review my thesis. I also sincerely thank Laurent Girin, Hervé Bredin and Nancy Bertin for being on my committee.

I would like to thank all the team members at Technicolor and Télécom ParisTech - especially the PhDs and post-docs, whom I interacted with the most, for making every day enjoyable. Special thanks to Thierry and Marlène for help with research data acquisition/storage and administrative difficulties, respectively.

I am also indebted to my schools, undergraduate and master's university - friends, teachers and places alike for sowing the seed of exploration and learning.

Finally, it would not have been possible without the unconditional love and support from friends and family. I must specially acknowledge numerous interesting technical

and non-technical conversations with my brother Jayneel, regular well-being/health reminders from my parents and innumerable jovial moments with my little nephew Ridharv.

Contents

1 Introduction

1.1 Objective and motivation

Humans invariably perceive the world multimodally. From a piece of cloth to a car, from a conversation to a concert, each object and event around us is discovered through perceptual senses of vision, sound, touch, smell and taste. Among our many innate abilities is the one to associate and use information from these different modalities (or senses) to understand and interact with our surroundings (Smith and Gasser, 2005). Specifically, with regard to audition and vision, humans are adept at meaningfully integrating auditory and visual characteristics of many real-world objects, events and actions. Let us imagine ourselves in the position of a person trying to cross the street depicted in Fig. 1.1. Among other things, we would like our audio-visual (AV) sensory systems to identify and locate the vehicles in our vicinity, listen to our neighbour amid the background noise, recognize occluded events from audio or inaudible objects from their appearance. It is noteworthy how we carry out these and many other related tasks with ease. Taking inspiration from this remarkable human capability, our objective in this thesis is to design algorithms that enable machines to describe and extract such objects and events from videos using joint AV analysis.

Figure 1.1. Navigating through a busy town, as a pedestrian, a cyclist or a driver leverages our ability to conduct audio-visual analysis of complex, dynamic scenes.

Robust scene understanding, as in our previous example, is achieved through both similarity and differences between what the audio and visual signals capture. Similarity allows enhancement of one signal using the other and complementarity enables compensating for one in the presence of the other. To illustrate this, note that temporal variations in sound are also indicated through visual motion for certain sources. Indeed, humans subconsciously use the lip movements to "enhance" what someone says in noisy environments (Krishnan et al., 2014). On the other hand, object or event identification through audio (or vision) is not hindered by noise or changes in the other modality. These are interesting phenomena that we wish to utilize and highlight in this thesis.

Research on this topic is not only interesting intellectually but also with regard to its practical applicability. The ubiquity of AV data opens up several application areas where jointly analyzing audio and visual signals could help both humans and machines alike. A few of them are listed below:

- Joint processing can be utilized for film post–production and more generally video

content creation and manipulation in the entertainment industry. Specifically, audio-visual speech analysis could aid the painstaking process of dubbing, mixing or automated dialogue replacement (ADR). The latter involves re-recording dialogues in a studio after filming due to issues such as synchronization, dramatic effects, line corrections, *etc.* (Woodhall, 2011). More recently it has found an interesting application for virtual character animation (Zhou et al., 2018; Shlizerman et al., 2018).

- Video content explosion on the Internet is now common knowledge. This research will find significant use in structuring and indexing this data for efficient storage and better content-based retrieval.

- Surveillance cameras capturing both audio and visual signals are becoming increasingly popular. However, these recordings are often of low quality (Ramakrishnan, 2018). This presents the perfect use case for joint processing to improve speech intelligibility, detect anomalous events for *e.g.* gunshots or enhance signals. Application to biometrics is also well-known (Bredin and Chollet, 2007).

- Assisting hearing/sight impaired people by using joint models to automatically go from audio to video synthesis and vice versa.

- Such research should find immediate applications in assisting unimodal approaches in provably important tasks such as video object segmentation and audio source separation.

- Translating multimodal sensory capabilities to algorithms will be a boon to the area of robotics. Robust AV analysis would not only support machine understanding but also interaction with surroundings and people (Karras et al., 2017).

1.2 Challenges

Promising research directions often come with equally daunting challenges and the topic under consideration is no exception. It is noteworthy that the mechanism and location where AV signals are bound in the brain to create multimodal objects still remain unknown to neuroscientists (Atilgan et al., 2018). With particular regard to this thesis, we must (i) define AV objects/events to decide what we want to look for; (ii) hypothesize ways of associating the modalities to discover the said objects; and (iii) deal with limited annotations, noise and scale of datasets we wish to use for training and testing. Some of these main difficulties are explained below. We tackle several of them in this thesis.

What is an AV object? Providing a general definition for the so called AV object is not straight-forward (Kubovy and Schutz, 2010). A recent study (Bizley et al., 2016) defines them as *a perceptual construct which occurs when a constellation of stimulus features are bound within the brain.* Such a definition encompasses correlated AV signals like the link between a speaker's lip movement and voice. However, whether the definition takes into account objects like cars where such relations are not visible, is a question of interpretation. In this thesis, for all practical purposes, we simply focus on the class of sounds which have an identifiable visual source. These would typically fall into one of the following sub-categories: (1) objects such as musical instruments with unique distinguishable sound characteristics, (2) objects such as vehicles which produce multiple sounds possibly from different sub-parts such as wheels, engine, *etc.* and (3) sounds like air horns which are produced by multiple visual sources. We rely on underlying tasks, data annotations and/or our formulations to evade the ambiguity and subjectivity associated with our definition. Also, we avoid commonly encountered visual objects such as television, tape recorder as these devices essentially re–produce sound and only add to the complexity of source identifiability and differentiability.

Intra–class audio and visual variations. Audio and visual characteristics of objects (events) can vary widely even within a single class. For a class like motorcycle the instance-level characteristics could be as diverse as the number of bike models. Other common sources of visual diversity are illumination, lighting, viewpoint and color variations.

Scarcity of annotated datasets. We wish to use videos from real-world events. While large amounts of such data can be obtained from websites like YouTube, its annotation is usually limited to a few, usually unreliable/noisy textual tags as provided by the user at upload time. This requires developing algorithms capable of learning and performing several complex tasks at scale with minimal supervision.

(a) Poor quality video where the audio event 'screaming' is not actually seen on–screen.

(b) Video modified to include irrelevant static frames

Figure 1.2. Some typical noise examples from user–generated YouTube videos.

Noise in data. Amateurish quality and unconstrained nature of user–generated videos result in different types of noise in both the modalities. Standard sources of noise include background audio and poor microphone or camera quality. Many videos are artificially altered to suit the content generator's needs. Specifically, the original audio track could be overlaid or completely replaced with music. In other cases, the content is modified or created with static, sometimes irrelevant images. Moreover,

the appearance of salient cues in both the modalities, like the visual appearance of a car and the sound of its engine may be asynchronous. This is to say that audio and visual cues appear at different times in the video. This is also referred to as the problem of alignment in some other works using visual and textual modalities (Alayrac et al., 2016). Some of these are shown in Fig. 1.2.

1.3 Contributions and outline

Having specified our focus on sounds produced by visual sources in the previous section, we are interested in investigating the following aspects of the problem:

1. Exploring AV fusion strategies;
2. Learning representations and analyzing their usefulness for scene understanding tasks, for instance audio source separation, visual object localization, *etc.*
3. Identifying the AV object.

In practice, audio and visual object extraction can be understood as performing audio source separation and spatio-temporal image/video segmentation, respectively.

While several factors could influence AV perception in humans, for this study we only have at our disposal generic videos acquired using single microphone and camera, as seen in Fig. 1.3. In such a scenario, two AV association mechanisms naturally emerge. The first, termed *motion–audio correlation* is motivated by the following observation: when playing an instrument or scratching a surface, motion used to excite objects and surfaces often dictates the temporal variations in sound. We use this fact to propose two novel non-negative matrix factorization (NMF) based motion-assisted audio source separation methods. Specifically, our initial approach couples sound–producing motion with audio through a soft ℓ_1 constraint. This provides encouraging separation results over a challenging multimodal music instrument dataset. To our best knowledge, this was the first study to use motion-capture data for audio source

separation. We overcome some limitations of this work, including application to videos with our next approach called *cross-modal regression*. Herein, we regress motion from temporal variations in audio within the NMF framework. The method gives as a by-product the ability to visually localize the sound–producing motion. Both the foregoing methods are shown to be particularly useful in challenging scenarios where multiple acoustically similar objects like violins in string quartets, "move" differently.

While such correlations have proved to be useful for source disambiguation, they are not explicitly visible in a variety of commonly encountered objects such as automobiles, washing machines, *etc.* They are non-existent for digital sounds emitted by objects like mobile phones. To tackle this case, we exploit *appearance–audio co–occurence*. This refers to associating general appearance (shape and/or color) of objects with their sounds based on co-occurrence in large datasets. To this end, we propose a weakly supervised deep representation learning framework based on multiple instance learning (MIL) to perform several tasks such as event classification, visual object localization and temporal localization of audio events simultaneously. The basic idea is to generate audio and visual proposals that may potentially contain the cues of interest. These are then scored according to relevance for a particular class. Building upon this work, we suggest a new audio proposal design to incorporate audio source separation capability and make classification more robust to interfering sounds.

1.3.1 Publications

The work discussed in this thesis has been a part of the following book chapter, patent, conference and workshop publications:

Book Chapter

- Slim Essid, Sanjeel Parekh, Ngoc Duong, Romain Serizel, Alexey Ozerov, Fabio Antonacci and Augusto Sarti, Multiview approaches to event detection and scene analysis. In T. Virtanen, M. Plumbley and D. Ellis (Eds.), Computational Ana-

lysis of Sound Scenes and Events, Springer 2018.

Conference and Workshop Papers

- Sanjeel Parekh, Alexey Ozerov, Slim Essid, Ngoc Duong, Patrick Pérez and Gaël Richard. Identify, Locate and Separate: Audio–visual object extraction in large video collections using weak supervision. IEEE WASPAA 2019 [To be submitted]
- Sanjeel Parekh, Slim Essid, Alexey Ozerov, Ngoc Duong, Patrick Pérez and Gaël Richard. Weakly supervised representation learning for unsynchronized audio–visual events, 2018. Extended abstract in CVPR Workshop on Sight and Sound 2018.
- Sanjeel Parekh, Slim Essid, Alexey Ozerov, Ngoc Duong, Patrick Pérez and Gaël Richard. Guiding audio source separation by video object information. IEEE WASPAA 2017.
- Sanjeel Parekh, Slim Essid, Alexey Ozerov, Ngoc Duong, Patrick Pérez and Gaël Richard. Motion informed audio source separation. IEEE ICASSP 2017.

Patents

- Sanjeel Parekh, Slim Essid, Alexey Ozerov, Ngoc Duong, Patrick Pérez and Gaël Richard. Weakly supervised learning for unsynchronized audio-visual events. Patent Application Filed, 2018.
- Sanjeel Parekh, Slim Essid, Alexey Ozerov, Ngoc Duong, Patrick Pérez and Gaël Richard. New approaches to motion informed audio source separation. US Patent Application 15956021, 2018.

1.3.2 Outline

In chapter 2 we review literature on audio-visual fusion techniques. This is followed by a discussion on ideas relevant to this thesis, namely source separation, non–negative matrix factorization and weakly supervised learning.

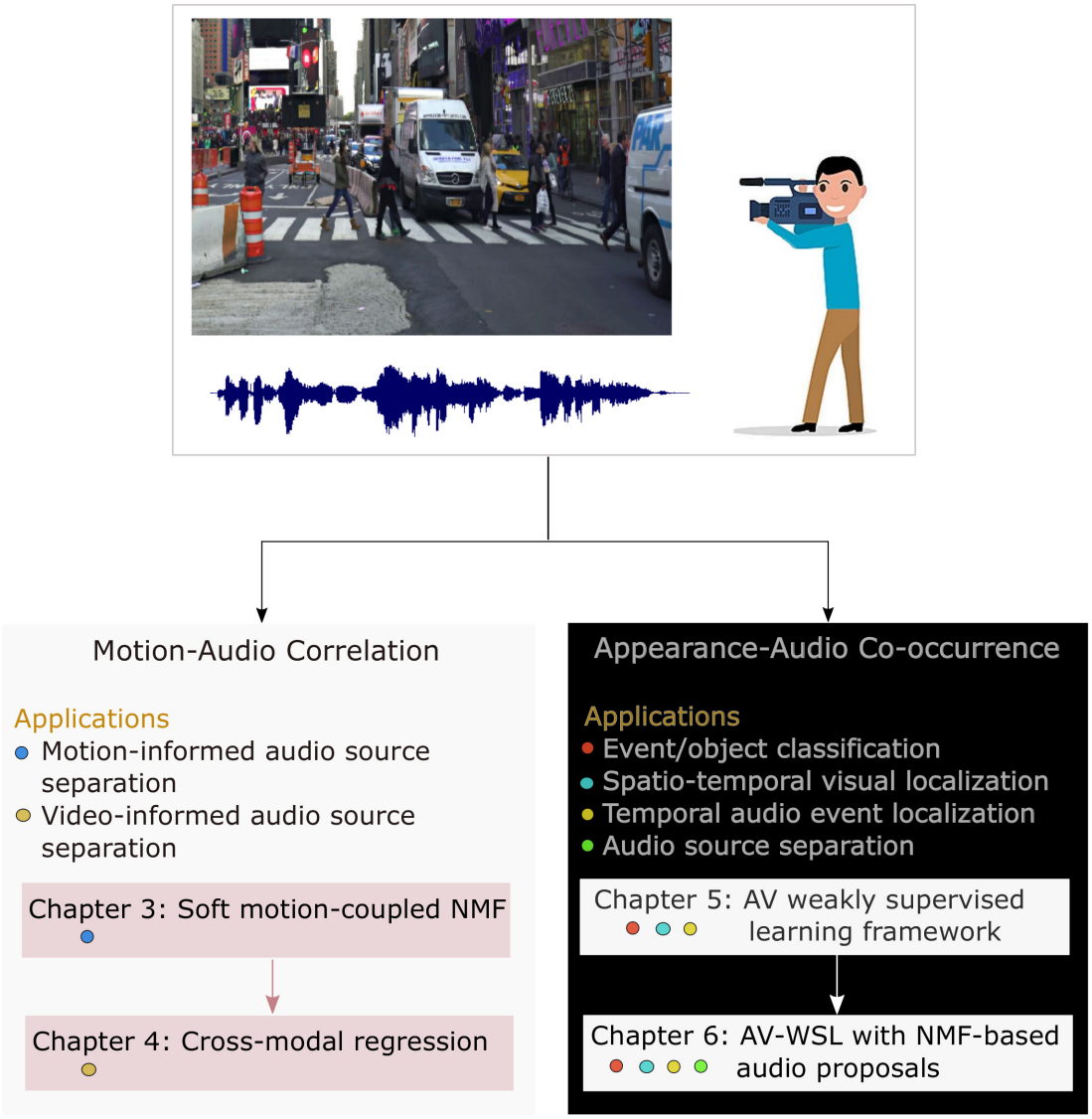


Figure 1.3. Thesis overview: The arrows between different chapters indicate dependencies. The colored dots under chapter 5 and 6 correspond to application bullet points and denote the ones we tackle in each.

In chapters 3 and 4 we detail soft motion coupled NMF and cross–modal regression, respectively. These are novel audio source separation algorithms utilizing motion–audio correlations.

In chapter 5 we put forth our weakly supervised deep learning framework for event classification and audio/visual cue localization. This is based on appearance audio co–occurrence in large datasets. We make the event/object classification more robust and incorporate source separation capabilities through a novel use of NMF for audio proposal design in chapter 6.

In chapter 7, we provide a summary of our contributions and future research perspectives, while throwing light on some recent developments in this area of multimodal analysis.

2

Related works and background

A generic pipeline to tackle different problems through joint audio-visual analysis typically involves early/intermediate or late fusion of extracted modality–specific representations and decisions, respectively. This simplified view is depicted in Fig. 2.1. Referring the reader to comprehensive review texts on commonly used hand–crafted/learnt audio (Serizel et al., 2018) and visual (Maragos et al., 2008) features, in this chapter we focus on important multimodal fusion strategies related to this thesis. Towards the end, we also briefly describe some relevant ideas that are repeatedly alluded to in the upcoming chapters.

2.1 Techniques for joint audio–visual analysis

2.1.1 Feature-space transformation methods

A number of techniques have been suggested to map the observed feature vectors from two modalities to a low dimensional space where a *measure of "dependency"* between them can be computed. Let us assume the N observed feature vectors from two modalities, $\mathbf{y}_{1,n} \in \mathbb{R}^{J_1}$ and $\mathbf{y}_{2,n} \in \mathbb{R}^{J_2}$ $(n = 1, \ldots, N)$, are assembled column-wise in

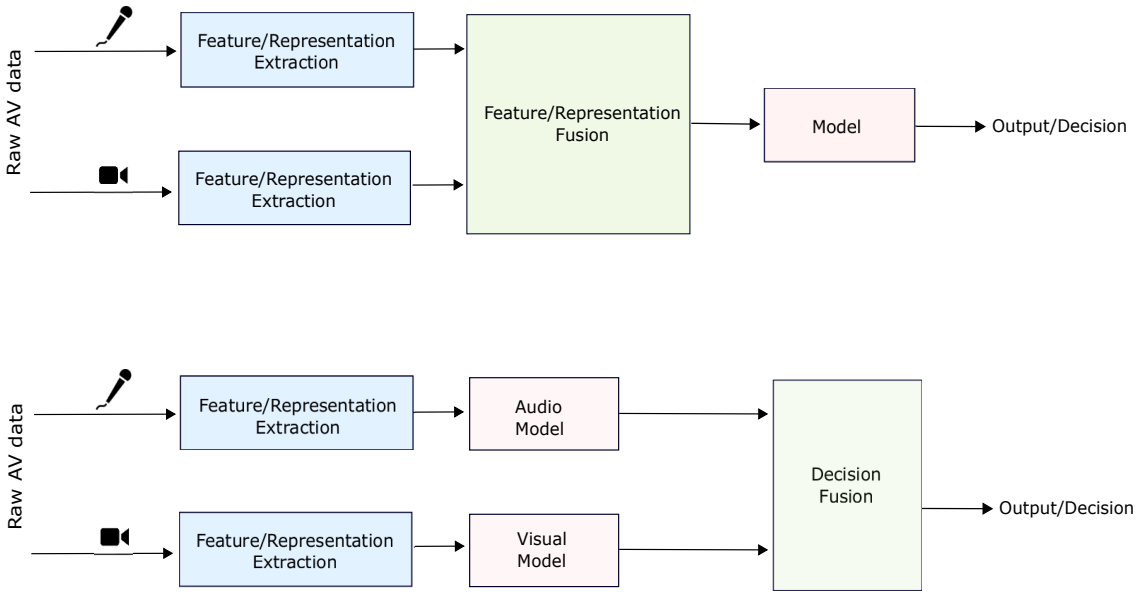


Figure 2.1. General system architectures for joint audio–visual analysis.

matrices $\mathbf{Y}_1 \in \mathbb{R}^{J_1 \times N}$ and $\mathbf{Y}_2 \in \mathbb{R}^{J_2 \times N}$ respectively.[1] The methods we describe here aim to find two mappings f_1 and f_2 (that reduce the dimensions of feature vectors in each modality), such that a dependency measure $S_{12}(f_1(\mathbf{Y}_1), f_2(\mathbf{Y}_2))$ is maximized. Various approaches can be described using this formalism. The advantages of doing so are two-fold: (i) it appropriately modifies the feature spaces to uncover relationships between modalities specified by the measure of dependency and (ii) by projecting data into the same space, dimensionality difference is eliminated and direct comparison across modalities is made possible. Fisher et al. (2001) choose the mutual information as a dependency measure and seek single-layer perceptrons f_1 and f_2 projecting the audiovisual feature vectors to a 2-dimensional space *i.e.* f_1 and f_2 each map their input to a 1-D space. Other more popular approaches (for which closed-form solutions can be found) use linear mappings to project the feature streams:

[1] The underlying assumption is that the (synchronized) features from both modalities are extracted at the same rate. In the case of audio and visual modalities this is often obtained by downsampling the audio features or upsampling the video features, or by using temporal integration techniques (Joder et al., 2008).

- Canonical correlation analysis (CCA), first introduced by Hotelling (Hotelling, 1936), aims at finding pairs of unit-norm vectors t_1 and t_2 such that

$$(t_1, t_2) = \underset{(t_1,t_2)\in\mathbb{R}^{J_1}\times\mathbb{R}^{J_2}}{\operatorname{argmax}} \operatorname{corr}\left(t_1^\mathsf{T}\mathbf{Y}_1, t_2^\mathsf{T}\mathbf{Y}_2\right). \tag{2.1}$$

  CCA can be considered equivalent to mutual information maximization for the particular case where the underlying distributions are elliptically symmetric (Kay, 1992). Several variants have been proposed to incorporate sparsity and non-negativity into the optimization problem to resolve issues with interpretability and ill-posedness, respectively (Kidron et al., 2005; Sigg et al., 2007). In the context of multimodal neuronal data analysis, temporal kernel CCA (Bießmann et al., 2010) has been proposed to take into account the temporal dynamics.

- An alternative to the previous methods (expected to be more robust than CCA) is co-inertia analysis (CoIA). It consists in maximizing the covariance between the projected audio and visual features:

$$(t_1, t_2) = \underset{(t_1,t_2)\in\mathbb{R}^{J_1}\times\mathbb{R}^{J_2}}{\operatorname{argmax}} \operatorname{cov}\left(t_1^\mathsf{T}\mathbf{Y}_1, t_2^\mathsf{T}\mathbf{Y}_2\right). \tag{2.2}$$

  A possible reason for CoIA's stability is that it is a trade-off between CCA and PCA, thus it benefits from advantages of both (Bredin and Chollet, 2006).

- Yet another configuration known as cross-modal factor analysis (CFA), and found to be more robust than CCA in (Li et al., 2003), seeks two matrices $\mathbf{T}_1 \in \mathbb{R}^{J\times J_1}$ and $\mathbf{T}_2 \in \mathbb{R}^{J\times J_2}$ where J is the dimensionality of the shared space, such that

$$(\mathbf{T}_1, \mathbf{T}_2) = \underset{(\mathbf{T}_1,\mathbf{T}_2)}{\operatorname{argmax}} \left(1 - \|\mathbf{T}_1\mathbf{Y}_1 - \mathbf{T}_2\mathbf{Y}_2\|_F^2\right) = \underset{(\mathbf{T}_1,\mathbf{T}_2)}{\arg\min} \|\mathbf{T}_1\mathbf{Y}_1 - \mathbf{T}_2\mathbf{Y}_2\|_F^2, \tag{2.3}$$

  with $\mathbf{T}_1\mathbf{T}_1^\mathsf{T} = I$ and $\mathbf{T}_2\mathbf{T}_2^\mathsf{T} = I$. $\|\mathbf{V}\|_F$ denotes the Frobenius norm of matrix $\mathbf{V}$.

Note that all the previous techniques can be kernelized to study non-linear coupling between the modalities considered (see for instance (Hardoon et al., 2004)). The

interested reader is referred to (Hotelling, 1936; Hardoon et al., 2004) for further details on these techniques, and to (Goecke and Millar, 2003) for a comparative study.

These methods have been primarily applied to the task of AV object localization and extraction which refers to the problem of identifying sources visually and/or aurally. The general approach is to first associate the two modalities using one of the discussed transformation techniques. The parameters learned during the former step can then be utilized for object localization and segmentation in both modalities. In particular, various approaches have leveraged the audio modality to better perform this task with the central idea of associating visual motion and audio. Fisher et al. (2001) proposed to use joint statistical modeling to perform this task using mutual information. Izadinia et al. (2013) consider the problem of moving-sounding object segmentation, using CCA to correlate audio and visual features. The video features consisting of mean velocity and acceleration computed over spatio-temporal segments are correlated with audio. The magnitude of the learned video projection vector indicates the strength of association between corresponding video segments and the audio. Several other works have followed the same line of reasoning while using different video features to represent motion (Kidron et al., 2005; Sigg et al., 2007).

2.1.2 Co-factorization techniques

Matrix factorization techniques can be profitably used to extract meaningful representations for the data being analyzed. Owing to multimodal data, one may resort to so-called *co-factorization* techniques, that is techniques performing two (or more) factorizations in parallel, which are linked in a particular way. Because of the different nature of the modalities, this link is usually characterized through temporal dependencies between them.

Hereby we focus on non–negative versions of co-factorization which are naturally used as such features commonly appear in audio and image applications (please refer to section 2.5 for justification regarding their usefulness for audio). Assuming that appropriate nonnegative features have been extracted at the same rate from the two modalities being analyzed—say the audio and images of a video—so that two

observation matrices $\mathbf{V}_1 \in \mathbb{R}_+^{J_1 \times N}$ and $\mathbf{V}_2 \in \mathbb{R}_+^{J_2 \times N}$ are available [2], for the audio and visual data. One may seek a model $(\mathbf{W}_1, \mathbf{W}_2, \mathbf{H})$, such that:

$$\begin{cases} \mathbf{V}_1 \approx \mathbf{W}_1 \mathbf{H} \\ \mathbf{V}_2 \approx \mathbf{W}_2 \mathbf{H} \\ \mathbf{W}_1 \geq 0,\ \mathbf{W}_2 \geq 0,\ \mathbf{H} \geq 0\,. \end{cases} \tag{2.4}$$

Here $\mathbf{W}_1, \mathbf{W}_2$ refer to modality–specific basis vectors and $\mathbf{H}$ to their temporal activations which are same for both modalities. This is referred to as *hard co-factorization*, an approach that has been followed in a number of works (see *e.g.* (Fitzgerald et al., 2009; Yoo and Choi, 2011; Yokoya et al., 2012)). Clearly, this approach is limited in that it does not account for possible local discrepancies across the modalities. This happens for example when there is a mismatch between the audio and the images information, say because of a visual occlusion in video analysis scenarios. This motivates the *soft co-factorization* model of Seichepine et al. (2014a), which merely encourages the temporal activations corresponding to each modality to be close, as opposed to equal, according to:

$$\begin{cases} \mathbf{V}_1 \approx \mathbf{W}_1 \mathbf{H}_1 \\ \mathbf{V}_2 \approx \mathbf{W}_2 \mathbf{H}_2 \\ \mathbf{H}_1 \approx \mathbf{H}_2 \\ \mathbf{W}_1 \geq 0,\ \mathbf{W}_2 \geq 0,\ \mathbf{H}_1 \geq 0,\ \mathbf{H}_2 \geq 0. \end{cases} \tag{2.5}$$

The model (2.5) is estimated by solving the following optimization problem:

$$\begin{cases} \min\limits_{\boldsymbol{\vartheta}} C_c(\boldsymbol{\vartheta})\ ;\ \boldsymbol{\vartheta} \triangleq (\mathbf{W}_1, \mathbf{H}_1, \mathbf{W}_2, \mathbf{H}_2) \\ \mathbf{W}_1 \geq 0,\ \mathbf{W}_2 \geq 0,\ \mathbf{H}_1 \geq 0,\ \mathbf{H}_2 \geq 0\,; \end{cases} \tag{2.6}$$

[2] To simplify, we consider the case of two modalities, but clearly the methods described here can be straightforwardly generalized to more than two data views by considering the relevant pairwise associations.

$$C_c(\boldsymbol{\vartheta}) \triangleq D_1(\mathbf{V}_1 \,|\, \mathbf{W}_1\mathbf{H}_1) + \gamma D_2(\mathbf{V}_2 \,|\, \mathbf{W}_2\mathbf{H}_2) + \delta P\left(\mathbf{H}_1, \mathbf{H}_2\right) ; \tag{2.7}$$

where:

- $D_1(.\,|\,.)$ and $D_2(.\,|\,.)$ are the measures of fit respectively relating to the first and second modalities; note that they may be chosen to be different divergences, each well suited to the corresponding feature space;
- $P\left(.,.\right)$ is a penalty on the difference between (properly rescaled) activation values occurring at the same instant; they can be for instance the ℓ_1 or ℓ_2-norm of the difference between the rescaled activations;
- γ and δ are regularization parameters controlling, respectively, the relative importance of each modality and the coupling penalty.

The interested reader is referred to (Seichepine et al., 2014a) for more details on the algorithms.

2.1.3 Joint AV codebook learning

These methods have been quite popular for the task of multimedia concept classification.[3] In essence, each element of these multimodal codebooks captures some part of a salient AV event. Work on short-term audiovisual atoms (S-AVA) (Jiang et al., 2009a) aims to construct a codebook from multimodal atoms which are a concatenation of features extracted from tracked short-term visual-regions and audio. An example S-AVA is illustrated in Fig. 2.2. To tackle the problem of video concept classification, this codebook is built through multiple instance learning.

Following this work, AV *grouplets* (AVG) (Jiang and Loui, 2011) were proposed, where separate dictionaries are constructed from coarse audio and visual foreground/background separation. Subsequently, AVGs are formed based on the mixed-and-matched temporal correlations. For instance, an AVG could consist of frames where a basketball player is seen in the foreground with the audio of the crowd cheering in the background.

[3] Here the term "concept classification" refer to generic categorization in terms of scene, event, object or location (Jiang et al., 2009a).

Figure 2.2. Example of an S-AVA taken from (Jiang et al., 2009a). The cake region track and the background birthday music form a salient multimodal atom for this event.

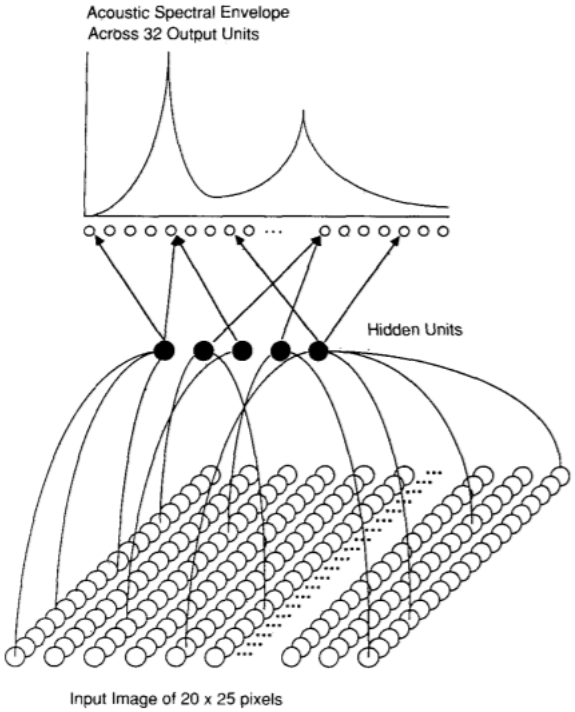
As an alternative, Jhuo et al. (2014) determine the relations between audio and visual modalities by constructing a bi-partite graph from their bag-of-words representation. Subsequently, spectral clustering is performed to partition and obtain *bi-modal* words. Unlike S-AVA and bimodal words, AVG has the advantage of explicitly tackling temporal interactions. However, like S-AVA, it relies on video region tracking, which is quite difficult for unconstrained videos.

2.1.4 Multimodal deep learning

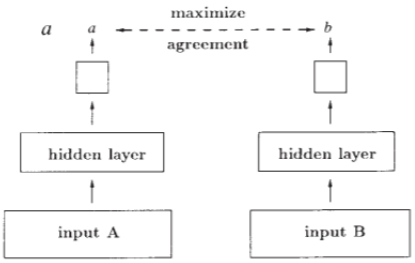
Lately, rapid progress in the application of deep learning methods to representation learning has motivated researchers to use them for fusing multimodal data.

The use of artificial neural networks for AV fusion can be traced back to Yuhas et al. (1989). The authors proposed to estimate acoustic spectral shapes from lip images for audio-visual speech recognition. In a later work, Becker and Hinton (1992) laid down the ideas for self organizing neural networks, where modules receiving separate but related inputs aim to produce similar outputs. Around the same time, ideas for self–supervised learning using different sensing modalities for classification of unlabeled data were put forth (de Sa, 1994). These foundational ideas are illustrated in Fig. 2.3.

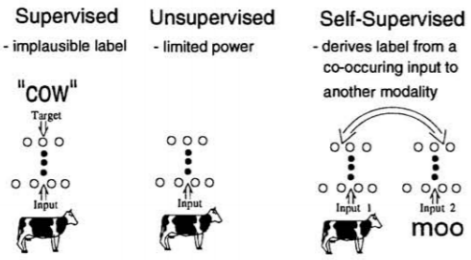
Owing to the advent of large scale datasets and training capabilities, each of these formulations have recently emerged in broader contexts for audio-visual fusion in generic videos. Notably, Owens et al. (2016a) train a convolutional neural network



(a) Illustration from (Yuhas et al., 1989) depicting acoustic spectral envelope estimation from lip images using ANN



(b) Learning by maximizing agreement between representations of two modalities. Figure from (Becker and Hinton, 1992)



(c) Self-supervised label using multimodal co-occurence. Illustration from (de Sa, 1994).

Figure 2.3. A depiction of early ideas for multimodal fusion and use of artificial neural networks (ANN).

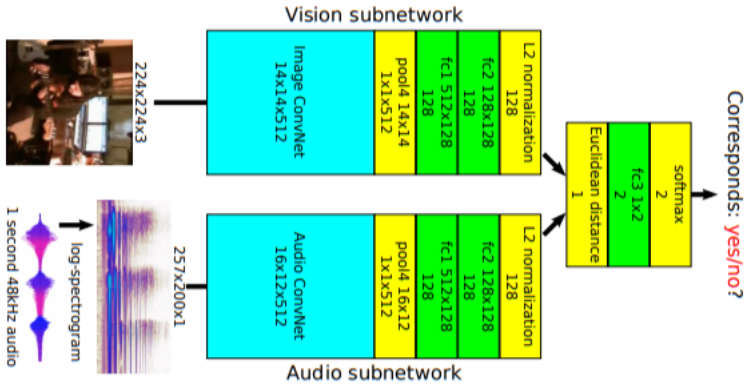


Figure 2.4. AV embedding network as depicted in (Arandjelović and Zisserman, 2017b)

(CNN) and recurrent neural network (RNN) based architecture to predict audio using visual input. In another work, this idea is extended to predict the audio category of static images (Owens et al., 2016b). Transfer learning experiments on image classification confirm that ambient audio assists visual learning (Owens et al., 2016b). This is reversed by (Aytar et al., 2016) to demonstrate how audio representation learning could be guided through visual object and scene understanding using a teacher-student learning framework. Herein the authors use trained visual networks to minimize the Kullback-Leibler divergence between the outputs of visual and audio networks. The resulting features are shown to be useful for audio event detection tasks. Subsequently, these ideas were extended to learning shared audio-visual-text representations (Aytar et al., 2017).

In some very recent works, useful AV representations are learnt through the auxiliary task of training a network to predict audio-visual correspondence (Arandjelović and Zisserman, 2017a,b). Indeed, the learnt audio representations in (Arandjelović and Zisserman, 2017a) achieve state-of-the-art results on audio event detection experiments. By design, such systems do not deal with the case of unsynchronized AV events as discussed earlier. Other notable approaches include multimodal autoencoder architectures (Ngiam et al., 2011) for learning shared representations even for the case where only a single view of data is present at training and testing time. Another interesting work extends CCA to learning two deep encodings, one for each view,

such that their correlation is maximized (Andrew et al., 2013). Interestingly, the structure of the network operations performed before the penultimate fully-connected layer in the architecture employed by (Arandjelović and Zisserman, 2017b), as seen in Fig. 2.4, is reminiscent of deep CCA. In particular, Kidron et al. (2005) showed that minimizing Euclidean distance between linearly transformed modality features with appropriate normalization is equivalent to maximizing their correlation. This is indeed the case for Arandjelović and Zisserman's system where the Euclidean distance between ℓ_2 normalized deep modality specific representations must be minimized for training.

Needless to say, the advent of deep learning, in particular end-to-end learning has blurred hard distinctions between feature extraction, fusion and model parameter learning stages (as seen in Fig. 2.1) and provided the flexibility to modify and model the inter–dependence between these.

2.2 A primer on some relevant problems and concepts

In this section we give a brief introduction to some relevant tasks and techniques that we repeatedly refer to or leverage in this thesis.

2.2.1 Source separation and non-negative matrix factorization

We directly delve into the problem of separating audio sources present in a single mixture. As in the case of AV objects, the definition of an audio source is application dependent. For a more general introduction to source separation we refer the reader to (Comon and Jutten, 2010; Vincent et al., 2018).

In its most general form the problem of single–channel audio source separation consists in obtaining an estimate for each of the J sources s_j forming the observed mixture $x(t)$ given by:

$$x(t) = f(s_1(t), s_2(t), \ldots, s_J(t)), \tag{2.8}$$

where f can be any linear or non-linear function and t denotes time indices. In this thesis, we assume that the mixing process is linear wherein the mixture $x(t)$ can be simply expressed as:

$$x(t) = \sum_{j=1}^{J} s_j(t). \tag{2.9}$$

Among several approaches developed to solve this under–determined problem, we mainly concern ourselves with a popular matrix decomposition method known as the non–negative matrix factorization (NMF) (Lee and Seung, 2001). Denoting by $\mathbf{X} \in \mathbb{C}^{F \times N}$ the mixture short–time Fourier transform (STFT) consisting of F frequency bins and N STFT frames, NMF can be used to decompose the magnitude $|\mathbf{X}|$ or power spectrogram $|\mathbf{X}|^2$, denoted by $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ such that,

$$\mathbf{V} \approx \mathbf{WH}, \tag{2.10}$$

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 composing the mixture $\mathbf{V}$ and their activation matrices, respectively. Here K is the total number of spectral patterns. Nonnegativity of the activation matrix $\mathbf{H}$ permits only additive combinations of the columns of $\mathbf{W}$. This encourages part-based decomposition which is essential for problems such as source separation. The idea is well depicted in Fig. 2.5 where we see how the time-frequency representation is broken down into constituent spectral templates and their temporal activation vectors. The NMF optimization problem to find suitable matrices $(\mathbf{W}, \mathbf{H})$ can be written as:

$$\begin{aligned} &\text{minimize} && D(\mathbf{V}|\mathbf{WH}) \\ &\text{subject to} && \mathbf{W} \geq 0, \mathbf{H} \geq 0, \end{aligned} \tag{2.11}$$

where $D(\mathbf{V}|\mathbf{WH}) = \sum_f \sum_n d(\mathbf{V}_{fn}|[\mathbf{WH}]_{fn})$ is a separable measure of fit. Functions from the β–divergence family are a popular choice for the scalar cost function $d(x|y)$:

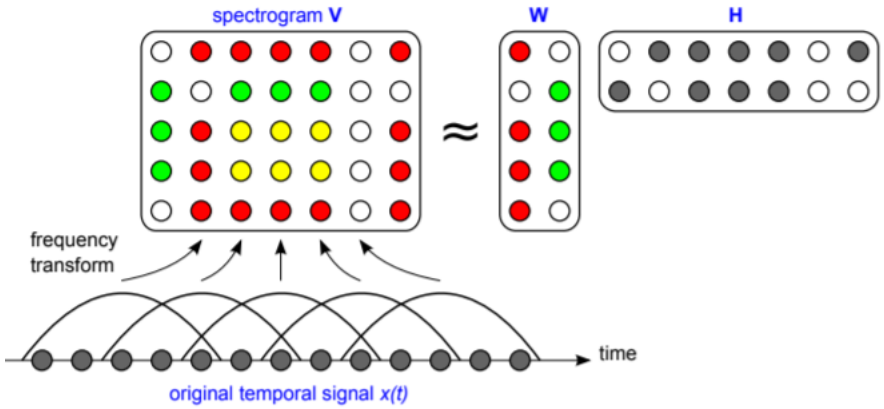


Figure 2.5. Schematic example of NMF applied to time-frequency audio representation. The blue and green columns depict unmixed source spectra that are correctly discovered as template vectors in $\mathbf{W}$ with $\mathbf{H}$ giving their mixing proportions. Illustration from (Févotte et al., 2018).

$$d(x|y) \triangleq \begin{cases} x \log \frac{x}{y} - x + y, & \beta = 1 \\ \frac{x}{y} - \log \frac{x}{y} - 1, & \beta = 0 \\ \frac{1}{\beta(\beta-1)}(x^\beta + (\beta-1)y^\beta - \beta x y^{(\beta-1)}), & \beta \in \mathbb{R} \setminus \{0, 1\}. \end{cases} \tag{2.12}$$

Here $\beta = 1$ and $\beta = 0$ correspond to popular Kullback-Leibler (KL) and Itakura-Saito (IS) divergence functions. Note that the cost function is jointly non–convex in $\mathbf{W}$ and $\mathbf{H}$. Iterative algorithms for estimating $\mathbf{W}$ and $\mathbf{H}$ commonly employ *multiplicative updates*. For the β–divergence these updates can be given by:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^\intercal((\mathbf{WH})^{(\beta-2)} \odot \mathbf{V})}{\mathbf{W}^\intercal(\mathbf{WH})^{(\beta-1)}} \tag{2.13}$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{((\mathbf{WH})^{(\beta-2)} \odot \mathbf{V})\mathbf{H}^\intercal}{(\mathbf{WH})^{(\beta-1)}\mathbf{H}^\intercal}. \tag{2.14}$$

Here products $\odot$, divisions $\frac{[.]}{[.]}$ and exponents are element-wise operations. $\mathbf{W}$ and $\mathbf{H}$ are initialized with random non–negative matrices. NMF is sensitive to initialization

and it is usually advised to run from multiple starting points (Févotte et al., 2018). Alternate initialization strategies, for instance using singular value decomposition (SVD) have also been considered (Boutsidis and Gallopoulos, 2008).

The updates given in equations (2.13) and (2.14) can be obtained using the following heuristic (Févotte and Idier, 2011): the derivative of the β–divergence with respect to a particular coefficient γ of $\mathbf{W}$ or $\mathbf{H}$ can be written as a difference of two nonnegative functions, $\nabla D(\gamma) = \nabla D^{+}(\gamma) - \nabla D^{-}(\gamma)$. The update can then be written as:

$$\gamma \leftarrow \gamma \frac{\nabla D^{-}(\gamma)}{\nabla D^{+}(\gamma)} \tag{2.15}$$

A more principled approach to arrive at a solution for such optimization problems is provided by the *majorization–minimization* technique (Hunter and Lange, 2004; Févotte and Idier, 2011). This requires construction and minimization of an auxiliary function that is an upper–bound of the original cost function, tight at the current value of the iterate ($\mathbf{W}$ or $\mathbf{H}$). By construction, this ensures that the objective function decreases in each iteration. Interestingly, the resulting updates are multiplicative and same as those given in equation for $\beta \in [1, 2]$ (Févotte and Idier, 2011).

For single channel audio source separation, after performing NMF decomposition one needs to group together specific components for source reconstruction. We mention below some standard audio-only methods of doing this:

- **Supervised NMF** (Wang and Plumbley, 2006): Herein, source–specific dictionaries, denoted by $\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_J$ are pre–computed from labeled training data. During test time, $\mathbf{H} = (\mathbf{H}_1^{\intercal}, \mathbf{H}_2^{\intercal}, \ldots, \mathbf{H}_J^{\intercal})^{\intercal}$ is estimated for any mixture by holding the learnt spectral patterns $\mathbf{W} = (\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_J)$ fixed, and applying the multiplicative update procedure given in equation 2.13. A variant called *semi–supervised NMF* (Mysore and Smaragdis, 2011) assumes that only a subset of source dictionaries are pre–learnt. The remaining are estimated together with source activation matrices $\mathbf{H}$ during test time.
- **Unsupervised NMF** (Spiertz and Gnann, 2009): This method involves no training phase. After performing NMF decomposition on a mixture, the basis

vectors are grouped into sources by performing clustering over their Mel–spectra. We discuss other unsupervised strategies in the forthcoming chapters.

Once the components are grouped into source basis vector and activation sub–matrices $\mathbf{W}_j$ and $\mathbf{H}_j$, each source STFT can be reconstructed using soft masking as:

$$\mathbf{S}_j = \frac{\mathbf{W}_j\mathbf{H}_j}{\mathbf{WH}} \odot \mathbf{X} \tag{2.16}$$

When considering power spectrograms for filtering, the above equation is equivalent to standard Wiener filtering. Lately, researchers have found soft masking with amplitude spectrograms to work better in practice (Liutkus and Badeau, 2015). Some theoretical justification has also been provided for such generalized soft masking (Liutkus and Badeau, 2015). The time domain signal can be obtained through inverse STFT operation on $\mathbf{S}_j$.

2.2.2 Weakly supervised learning

As already discussed in the previous chapter, fully annotated training data for AV tasks is seldom available. This has led to use and development of alternative non-supervised learning approaches such as reinforcement learning, self-supervised and weakly supervised learning. In this thesis, we primarily focus on weakly supervised learning which is an umbrella term used to refer to training strategies that only have at their disposal incomplete, possibly noisy information with regard to one or more tasks. For instance, it is often easier to simply indicate the presence or absence of an object in a large number of images. This paradigm allows precisely localizing the said object in these images with only such binary information.

Multiple instance learning (MIL) (Dietterich et al., 1997a) is a popular framework to achieve this. Different from supervised learning, herein, the data is given as bags with associated labels, where each bag could possibly contain multiple instances. In our running example, each image could be considered a bag, with different regions as the constituent instances. In addition to building a model for bag classification, an important goal of MIL approaches is to discover the key instances that lead to

the bag label. This is extremely important for interpretability. Several problems in bioinformatics (Kraus et al., 2016), medical imaging (Ilse et al., 2018), computer vision (Oquab et al., 2015) have been adapted to fit such a formulation. This is also key to our system discussed in chapters 5 and 6.

Broadly speaking, MIL approaches can be categorized into bag-space and instance-space methods (Amores, 2013). Bag space methods usually aim to solve the bag–level classification problem either by using supervised classification techniques over a global bag-level embedding vector (Dong, 2006) or through distance–based classifiers such as k-NN (Wang and Zucker, 2000). On the other hand, instance–space methods aggregate local information from instance–level classifiers or features to train models for bag–level classification. This has the added advantage of revealing key instances responsible for classification. Several support vector machine (SVM) and neural network based methods (Ilse et al., 2018; Oquab et al., 2015) have taken this path. We defer the discussion of relevant MIL literature for computer vision and audition to chapter 5.

3

Motion-informed audio source separation

Synopsis

In this chapter we tackle the problem of single channel audio source separation driven by descriptors of the sounding object's motion. As opposed to previous approaches, motion is included as a soft coupling constraint within the non–negative matrix factorization framework. The proposed method is applied to a multimodal dataset of instruments in string quartet performance recordings where bow motion information is used for separation of string instruments. We show that the approach offers better source separation results than an audio-based baseline and the state-of-the-art multimodal-based approaches on these very challenging music mixtures.

3.1 Introduction

Consider the scene of a busy street or a music concert: what we hear in these scenarios is a mix of sounds coming from multiple sources. However, information received from the human visual apparatus in terms of movement of these sources over time is very useful for decomposing and associating them with their respective audio streams (Chen et al., 2002). Indeed, when performing any action or interacting

with sound-producing objects such as musical instruments, motion often dictates the temporal evolution of the produced sounds. In this chapter, we are interested in using this correlation between audio and motion to perform the challenging task of single channel audio source separation.

Several approaches have been proposed for monaural source separation in the unimodal case, *i.e.*, methods using only audio (Wang and Plumbley, 2006; Durrieu et al., 2011; Huang et al., 2014; Févotte et al., 2018), in which NMF has been among the most popular ones. Typically, source separation in the NMF framework is performed in a supervised manner (Wang and Plumbley, 2006), where the magnitude or power spectrogram of an audio mixture is factorized into nonegative spectral patterns and their activations. In the training phase, spectral patterns are learnt over clean source examples and then factorization is performed over test examples while keeping the learnt spectral patterns fixed. In the last few years, several methods have been proposed to group together appropriate spectral patterns for source estimation without the need for the above mentioned dictionary learning step. Spiertz and Gnann (2009) proposed a promising generic basis vector clustering approach using Mel-spectra (Davis and Mermelstein, 1990). Subsequently, methods based on shifted-NMF, inspired by western music theory and linear predictive coding were proposed (Jaiswal et al., 2011; Guo et al., 2015). While the latter has been shown to work well with harmonic sounds, its applicability to percussive sounds will be limited.

In the single channel case, it is possible to improve system performance and avoid the spectral pattern learning phase by incorporating auxiliary information about the sources. The inclusion of side information to guide source separation has been explored within task-specific scenarios such as text informed separation for speech (Le Magoarou et al., 2015) or score-informed separation for classical music (Fritsch and Plumbley, 2013). Recently, there has also been much interest in user-assisted source separation where the side information is obtained by asking the user to hum, speak or provide time-frequency annotations (Smaragdis and Mysore, 2009; Liutkus et al., 2013; Duong et al., 2014). For brevity, here we only elaborate on methods utilizing motion data as auxiliary information.

In most cases, information about motion is extracted from the video images. Early

work by Fisher et al. (2001) sought to learn a multimodal embedding through mutual information (MI) maximization. This is then used to tackle the task of user-assisted audio enhancement. However, the method used for computing MI *i.e.* Parzen window estimation is complex and may suffer in quality when the data used to perform the estimation is limited. A different technique that aims to extract AV independent components (Smaragdis and Casey, 2003) does not work well with dynamic scenes containing moving objects. Later, work by Barzelay and Schechner (2007) considered onset coincidence to identify AV objects and subsequently perform source separation. They delineate several limitations of their work, including: setting multiple parameters for optimal performance on each example and possible performance degradation in dense audio environments. AV source separation has also been attempted using sparse representations (Casanovas et al., 2010). However, the method's application is limited due to its dependence on active-alone regions to learn source characteristics. It is also assumed that all the audio sources are seen on-screen, which is not always realistic. A recent work proposes to perform AV source separation and association for music videos using music score information (Li et al., 2016a). Some prior work on AV speech separation has also been carried out (Nakadai et al., 2002; Rivet et al., 2007), primary drawbacks being the large number of parameters and hardware requirements.

In this work we improve upon several limitations of earlier methods. Our approach utilizes motion information within the NMF parameter estimation procedure through soft coupling rather than a separate step after factorization. This not only preserves flexibility and efficiency of the NMF system, but unlike previous motion-based approaches, significantly reduces the number of parameters to tune for optimal performance (to effectively just one). Particularly, we show that in highly non-stationary scenarios, information from motion related to the causes of sound vibration from each source can be very useful for source separation. With the exception of a recently published study (Sedighin et al., 2016), to the best of our knowledge no previous work has incorporated motion into the NMF-based source separation systems. Moreover, as we demonstrate in Section 3.3, the applicability of methods proposed in (Sedighin et al., 2016) is limited. Our method is applied to musical instrument source separation in string trios using bow motion information. To the best of our knowledge this chapter

describes the first study to use motion capture data for audio source separation.

Chapter outline. We begin by describing the problem and the proposed co-factorization approach tying temporal evolution of audio to motion in section 3.2. Experimental results for the task of source separation are presented in section 3.3 while limitations and future work are discussed in section 3.4.

3.2 Soft motion coupled NMF

3.2.1 Problem formulation

As already discussed in section 2.2.1, the problem of single channel audio source separation consists in obtaining an estimate for each of the J sources s_j forming the observed linear mixture $x(t)$ (equation 2.9).

Using NMF we can decompose the mixture magnitude or power spectrogram $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ consisting of F frequency bins and N short-time Fourier transform (STFT) frames, into two nonnegative matrices $\mathbf{W} \in \mathbb{R}_+^{F \times K}$and $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ interpreted as the nonnegative audio spectral patterns and their activation matrices, respectively

Additionally, here we assume that information about the causes of sound vibration of each source is known in the form of motion activation matrices $\mathbf{H}_{m_j} \in \mathbb{R}_+^{K_{m_j} \times N}$, vertically stacked into a matrix $\mathbf{H}_m \in \mathbb{R}_+^{K_m \times N}$:

$$\mathbf{H}_m = \begin{bmatrix} \mathbf{H}_{m_1} \\ \vdots \\ \mathbf{H}_{m_J} \end{bmatrix}, \text{ where } K_m = \sum_{j=1}^{J} K_{m_j}. \tag{3.1}$$

Thus, our objective is to cluster the K spectral patterns into J audio sources using the motion activation matrix $\mathbf{H}_m$.

3.2.2 Approach

Following Seichepine et al. (2014b), our central idea is to couple $\mathbf{H}_m$ with the audio activations, *i.e.*, to factorize $\mathbf{V}$ such that $\mathbf{H}$ is "similar" to $\mathbf{H}_m$. With such a constraint, the audio activations for each source $\mathbf{H}_j$ would automatically be coupled with their counterparts in the motion modality $\mathbf{H}_{m_j}$ and we would obtain basis vectors clustered into audio sources. For this purpose, we propose to solve the following optimization problem with respect to $\mathbf{W}, \mathbf{H}$ and $\mathbf{S}$:

$$\begin{aligned} &\underset{\mathbf{W},\mathbf{H},\mathbf{S}}{\text{minimize}} \left[D_{KL}(\mathbf{V}|\mathbf{WH}) + \alpha\|\Lambda_a\mathbf{H} - \mathbf{SH}_m\|_1 + \beta \sum_{k=1}^{K}\sum_{n=2}^{N}(\lambda_{a,k}h_{kn} - \lambda_{a,k}h_{k(n-1)})^2 \right] \\ &\text{subject to } \mathbf{W} \geq 0, \mathbf{H} \geq 0. \end{aligned} \tag{3.2}$$

In equation (3.2), the first term is the standard generalized Kullback-Leibler (KL) divergence cost function (equation 2.12). The second term enforces "similarity" between audio and motion activations, up to a scaling diagonal matrix $\mathbf{S}$, by penalizing their difference with the ℓ_1 norm. The last term is introduced to ensure ℓ_2 temporal smoothness of the audio activations (Virtanen, 2007). The influence of each of the last two terms on the overall cost function is controlled by the hyperparameters α and β, repectively. Λ_a is a diagonal matrix with k^{th} diagonal coefficient given by $\lambda_{a,k} = \sum_f w_{fk}$ which simply scales $\mathbf{H}$ (or equivalently, normalizes the columns of $\mathbf{W}$ (Essid and Févotte, 2013)).

The cost function is minimized using a block coordinate majorization-minimization (MM) algorithm where $\mathbf{W}$ and $\mathbf{H}$ are updated sequentially (Seichepine et al., 2014b). Our formulation is a simplified variant of the previously proposed soft non-negative matrix co-factorization (sNMcF) algorithm (Seichepine et al., 2014b), wherein two modalities are factorized jointly with a penalty term soft-coupling their activations. However, here we do not factorize the second modality (*i.e.*, the motion modality) and its activations are held constant in the update procedure. Note that, from the model's perspective, $\mathbf{H}$ and $\mathbf{H}_m$ need not contain the same number of components.

So if $K \neq K_m$, then we can readily ignore some components when coupling. However, for this work we maintain $K = K_m$. The reader is referred to (Seichepine et al., 2014b) for details about the algorithm. Reconstruction is done by performing soft masking as detailed in equation 2.16.

In the following section, we will discuss the procedure for obtaining motion activation matrices $\mathbf{H}_{m_j}$ for each source.

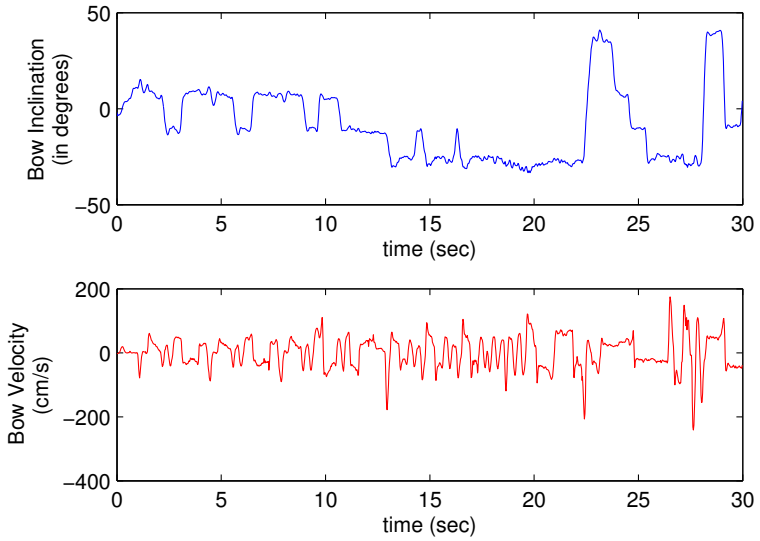


Figure 3.1. An example of bow inclination and velocity data for violin from the multimodal EEP dataset (Marchini, 2014).

3.2.3 Motion modality representation

While for audio, the classic magnitude spectrogram representation is used, motion information must be processed to obtain a representation that can be coupled with audio activations. The question now being: What motion features will be useful?

We work with a multimodal dataset of instruments in string quartet performance recordings. In this dataset (see section 3.3.1), the motion information exists in the form of tracking data (motion capture or MoCap data) acquired by sensors placed

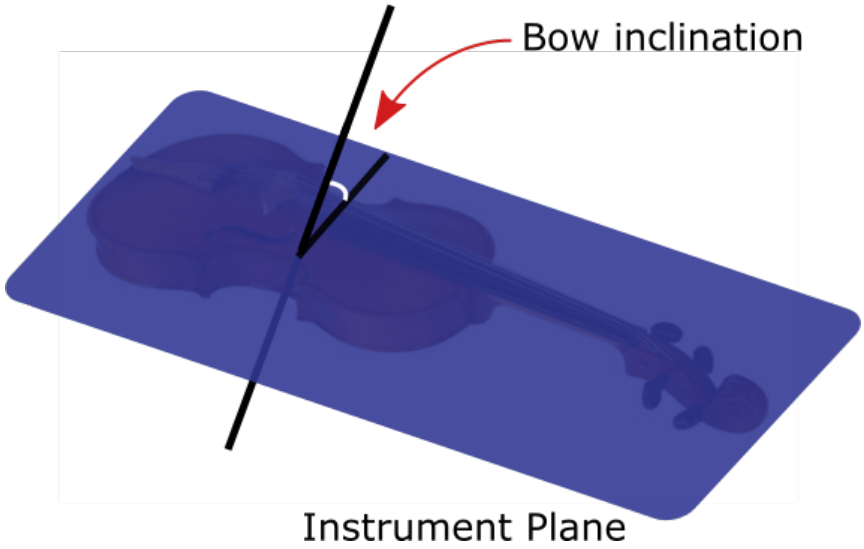
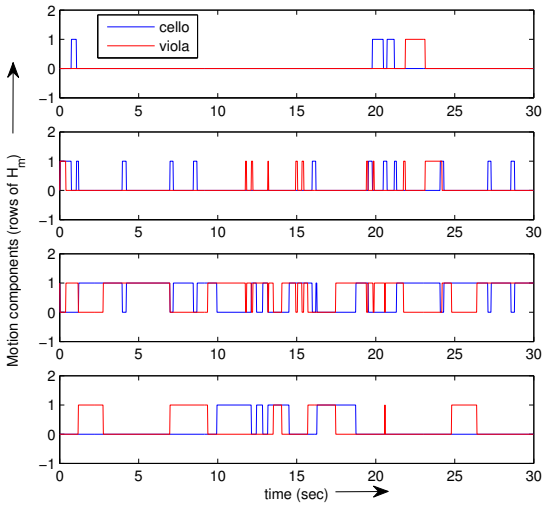


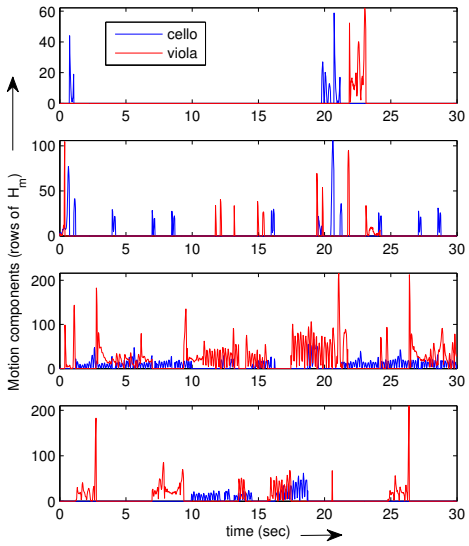
Figure 3.2. A visualization for bow inclination: the angle between the instrument plane and the bow.

on each instrument and the bow (Marchini, 2014). Now we immediately recognize that information about "where" and "how" strongly the sound-producing object is excited will be readily conveyed by bowing motion velocity and orientation in time. In this light, we choose to use bow inclination (in degrees) and bow velocity (cm/s) as features (as shown in Fig. 3.1), which can be easily computed from the raw motion capture data described in (Maestre, 2009; Marchini, 2014). These descriptors have been pre-computed and provided with the dataset. The bow inclination is defined as the angle between the instrument plane and the bow, as depicted in Fig. 3.2. The bow velocity is the time derivative of the bow transversal position. We refer the reader to Maestre's thesis work (Maestre, 2009) for a mathematical definition of these quantities. The motion activation matrix, $\mathbf{H}_{m_j}$ for $j \in (1, J)$ is then built using the following simple strategy:

1. In the first step, we uniformly quantize the bow inclination for each instrument into 4 bins based on the maximum and minimum inclination values. A binary encoded matrix of size $4 \times N$ is then created where the row corresponding to the active bin is set to 1 and the rest to 0 for each time frame.



(a) Quantized bow inclination.



(b) Quantized components multiplied with bow velocity.

Figure 3.3. Motion representation.

2. With such a simple descriptor we already have information about the active string within each time window. We then do a pointwise multiplication of each component with the absolute value of the bow velocity. Intuitively, this gives us information about string excitation. Fig. 3.3 visualizes the effectiveness of this step, where Fig. 3.3a depicts the quantized bow inclination vector components, overlapped for two sources. Notice, especially in the third subplot, that there are several places where the components overlap and the contrast between the motion of these sources is difficult to see. However, once it is multiplied with the bow velocity (in Fig. 3.3b) the differences are much more visible.

3.3 Experimental validation

We conduct several tests over a set of challenging mixtures to judge the performance of the proposed approach.

3.3.1 Dataset

We use the publicly available Ensemble Expressive Performance (EEP) dataset[1] (Marchini et al., 2014). This dataset contains 23 multimodal recordings of string quartet performances (including both ensemble and solo). These recordings are divided into 5 excerpts from Beethoven's Concerto N.4, Op. 18. Four of these, labeled from P1 to P4 contain solo performances, where each instrument plays its own part in the piece. We use these solo recordings to create mixtures for source separation. Note that due to unavailability of microphone recording for the solo performance of the second violin in the quartet we consider mixtures of three sources, namely: Violin (vln), Viola (vla) and Cello (cel). The acquired multimodal data consists of audio tracks and motion capture for each musician's instrument performance.

[1] http://mtg.upf.edu/download/datasets/eep-dataset

3.3.2 Experimental setup

For evaluating the performance of the proposed methods in different scenarios we consider the following three different mixture sets:

1. **Set 1** – 4 trios of violin, viola and cello, one for each piece denoted by P1, P2, P3, P4 in Table 3.1.
2. **Set 2** – 6 two-source combinations of the three instruments for pieces P1 and P2.
3. **Set 3** – 3 two-source combinations of the same instrument from different pieces, *e.g.*, a mix of 2 violins from P1 and P2.

Our approach is compared with the following baseline and state-of-the-art methods:

1. **Mel NMF** (Spiertz and Gnann, 2009) – This is a unimodal approach where basis vectors learned from the mixture are clustered based on the similarity of their Mel-spectra. We take help of the example code provided online for implementation of this baseline method.[2]
2. **MM Initialization** (Sedighin et al., 2016) – This is a multimodal method where the audio activation matrix is initialized with the motion activation matrix during the NMF parameter estimation.
3. **MM Clustering** (Sedighin et al., 2016) – Here, after performing NMF on audio, basis vectors are clustered based on the similarity between motion and audio activations. The basic idea behind the clustering algorithm is that if an audio activation vector, $\mathbf{h}_k$ belongs to source s_j, then the rank of the matrix obtained after stacking together $\mathbf{h}_k$ and bow velocity of the corresponding source should be close to 1. For more details the reader is referred to (Sedighin et al., 2016).

Note that, for the latter two methods, as done by the authors, we utilize the Itakura-Saito (IS) divergence cost function. Code provided by Févotte and Idier (2011) is used for standard NMF algorithms.

The audio is sampled at 44.1 kHz. We compute the spectrogram with a Hamming window of size 4096 (92 ms) and 75% overlap for each 30 sec excerpt. Thus, we have a $2049 \times N$ matrix. Here N is the number of STFT frames. Since the MoCap data

[2] http://www.ient.rwth-aachen.de/cms/dafx09/

is sampled at 240 Hz, each of the selected descriptors is resampled to match the N STFT audio frames. For all the runs the proposed method hyperparameters were set at $\alpha = 10$ and $\beta = 0.3$ after preliminary testing. As discussed in section 3.2.3, the number of components for each instrument is set to 4. NMF for each of the methods is run for 100 iterations. For each mixture, all the methods are run 5 times due to random initialization and the reconstruction is performed using a soft mask. The average of each evaluation metric over these runs is displayed in Table 3.1.

Evaluation metrics: We report the Signal to Distortion Ratio (SDR) computed using the BSS_EVAL Toolbox version 3.0 (Vincent et al., 2006). This standard quantitative performance metric for source separation is defined as:

$$\text{SDR} := 10 \log_{10} \frac{s_{\text{target}}}{e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}} \tag{3.3}$$

where s_{target} is the projection of estimated source onto the reference source signal and $e_{\text{interf}}, e_{\text{noise}}, e_{\text{artif}}$ are the interference, noise and artifact error terms, respectively.

3.3.3 Results and discussion

The results are as presented in Table 3.1[3], where the best SDR for each mixture is displayed in bold. Our method clearly outperforms the baselines and the state-of-the-art methods for highly challenging cases of trios (Set 1) and duos involving the same instrument (Set 3). For the third set of mixtures, where we compose mixtures of similar instruments, audio only methods would not be able to cluster the spectral patterns well as the timbral features would be very similar. Motion information clearly plays a crucial role for disambiguation and indeed the proposed method outperforms all the others by a large margin.

Particularly, notice that the multimodal baselines do not perform well. The MM

[3] Please note that in chapters 3 and 4 we fixed two issues with our STFT–ISTFT code: (1) corrected edge processing by zero-padding the input and (2) used a *periodic* window to make sum of overlapping windows constant for overlap-add scheme. These changes and random initialization lead to slight differences in values when compared with their published versions. However, our conclusions remain unchanged.

initialization relies on setting to zero the coefficients where there is no motion. This might not prove to be the best strategy with such a dataset because even during the inactive period of the audio there is some motion of the hand. On the other hand, multimodal clustering depends on the similarity between source motion activation centroids and audio activations. As we observe during the experiments, such a similarity is not very obvious for the data we use and the method ends up assigning most vectors to a particular cluster.[4]

Despite its overall good performance it is worth noting that for trio mixtures the proposed method performs poorly with P2. In fact, all the mixtures involving the viola from the second piece seem to have worse performance than others. We note that the separation for the viola suffers. One possible reason for this could be that, for P2, the motion descriptors of the viola with respect to the violin and the cello overlap in parts. As a consequence, the estimation of $\mathbf{W}$ for such cases is poor.

Note that the optimal value for α, which is held constant here, would differ for each recording. Thus, it should be possible to tune that parameter to gain the best performance, as could be achieved by an audio engineer through a knob controlling α, in a real world audio production setting. Also, note that we work with a limited number of components which is probably not well suited for some of these cases.

3.4 Conclusion

We have proposed a method for coupling motion and audio activations through soft ℓ_1 constraint within the NMF framework. This allows us to jointly factorize and cluster the audio spectral components. The results obtained on the multimodal string instrument dataset are encouraging and serve as a proof-of-concept for using motion as auxiliary information for NMF-based source separation. Moreover, motion is shown to be particularly useful for mixtures of acoustically similar sources that "move" differently. However, as we discuss next, this method's applicability to unconstrained

[4] We also observed that MM methods with KL divergence did not yield better results than IS.

Table 3.1. SDR (measured in dB) for different methods on each mixture. Best SDR is displayed in bold.

	Mixtures	Proposed Method	MM Init	MM Clustering	Mel NMF
Set 1	P1	**2.6**	−2.0	−7.2	−0.2
	P2	−0.5	−1.5	−7.6	**−0.1**
	P3	**1.3**	−0.4	−6.8	−1.2
	P4	**2.0**	−0.3	−7.4	−0.1
Set 2	P1 - vln + vla	**4.2**	0.6	0.2	0.6
	P1 - vln + cel	**7.0**	3.1	−3.8	2.8
	P1 - vla + cel	**2.4**	−1.3	−4.5	1.4
	P2 - vln + vla	0.0	−2.2	−1.2	**1.8**
	P2 - vln + cel	**5.8**	4.9	−3.6	4.8
	P2 - vla + cel	3.1	4.4	−3.7	**4.7**
Set 3	vln (P1) + vln (P2)	**3.5**	0.7	0.7	−0.1
	vla (P1) + vla (P2)	**−0.3**	−1.4	−3.9	−0.6
	cel (P1) + cel (P2)	**3.5**	1.7	−5.8	−1.8

videos where no explicit motion information is provided is not straightforward. This leads us to our second approach termed cross-modal regression.

4

Cross-modal regression for separating visually-indicated sounds

Synopsis

We introduce novel joint and sequential multimodal approaches for the task of single channel audio source separation in videos. Specifically, we present methods that utilize the non-negative least squares formulation to couple motion information extracted from video with audio. The proposed techniques generalize work discussed in the previous chapter and easily extend to video data. Experiments with two distinct multimodal datasets of string instrument performance recordings illustrate their advantages over the existing methods.

4.1 Introduction

Several sounds in the real world are *visually indicated* through their relation to the sound-producing motion. This chapter focuses on single channel audio source separation in audiovisual recordings of such sound mixtures by leveraging the accompanying motion information in the visual stream. While the previous chapter provided a proof-of-concept for coupling motion and audio, its straightforward application to

generic videos is difficult. In particular, we explicitly constructed a motion activation matrix using specific inputs such as bow inclination. It is difficult to obtain precise information about quantities such as inclination in real world scenarios. Moreover, the formulation makes it difficult to flexibly modify the number of audio basis vectors. These factors limit the method's applicability and performance. In this chapter, we extend and improve upon these shortcomings by proposing joint and sequential NMF-based approaches for linking motion and audio through cross-modal regression.

The intuition behind this work is as follows: motion features such as velocity, obtained from visual analysis, encode information about the physical excitation of a sounding object. On the other hand, for the audio modality, a representation of this excitation can be found in the spectral component activation matrix obtained after NMF decomposition. Thus, our hypothesis is that a set of audio activations would be "similar" to the velocity of *sound-producing* motion. We establish the idea's effectiveness for audio source separation through experiments on two very challenging multimodal string quartet performance datasets involving video and motion capture data.

Chapter outline. In section 4.2 we discuss various components of our approach such as motion processing and the proposed NMF parameter estimation strategies. This is followed by experimental results for source separation on two music performance datasets in section 4.3. We also show qualitative results for visual localization of the sound-producing motion, which is obtained as a by-product of our method.

4.2 Proposed approach

As already discussed, when dealing with only a single mixture, without a training step as in the supervised case (Wang and Plumbley, 2006), the source separation problem in the NMF framework reduces to assigning the appropriate spectral patterns, *i.e.*, each of the K components in the columns of $\mathbf{W}$, to the J sources. Here we propose methods to guide this assignment by using the associated source-specific

motion information through cross-modal regression. We discuss next each building block of our approach depicted in Fig. 4.1. We take as input the audio mixture and a set of video frames. The input frames are processed to extract motion features (section 4.2.1), while a time-frequency representation is computed from the audio mix. These are then used for estimating model parameters (section 4.2.2) and source reconstruction (section 4.2.3).

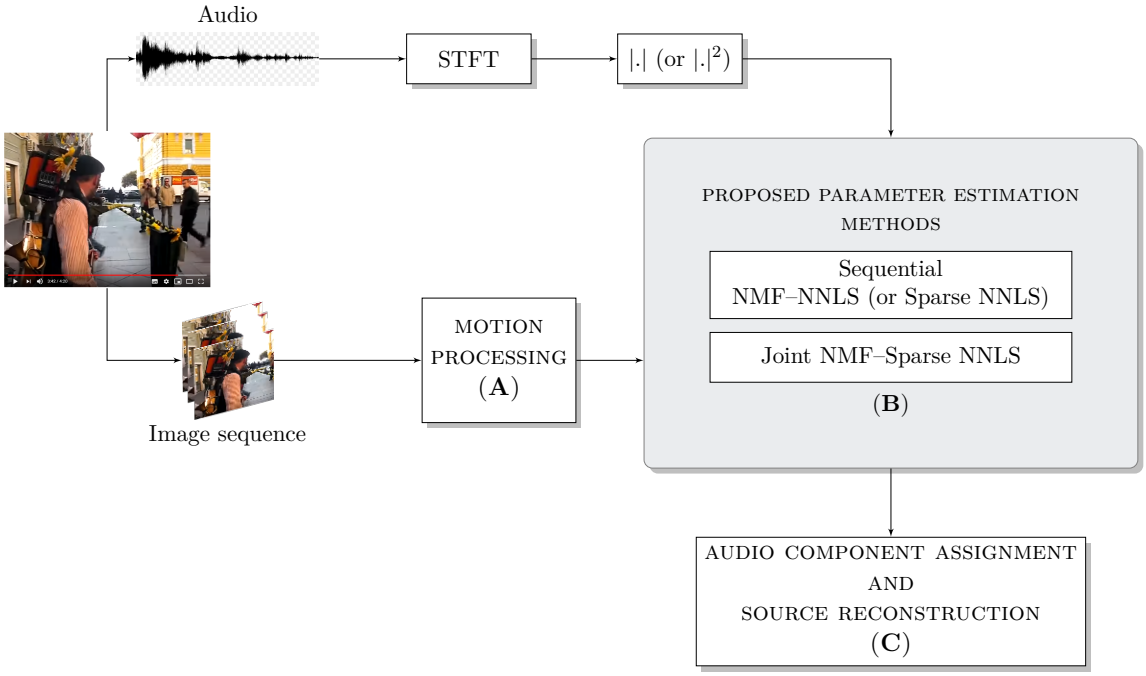


Figure 4.1. Overview of our approach. The image sequence is processed to extract average velocities over motion clusters (**A**). This information is used together with audio for clustering NMF spectral components in the parameter estimation block (**B**). Finally, we reconstruct each source based on the estimated clusters (**C**).

4.2.1 Motion processing unit

The motion data must be suitably processed to establish meaningful cross-modal relations. To begin with, for image sequences extracted from videos we assume that the spatial location of each moving AV object is known (in a user-assisted manner

or otherwise), as shown through the bounding boxes in Fig. 4.2. Note that J audio sources correspond to the same number of visual objects in the images. For each such moving region we proceed as follows:

- First, motion trajectory segmentation is performed on the image sequence using a state-of-the-art multicuts-based formulation (Keuper et al., 2015). As shown in Fig. 4.2, here the idea is to cluster point trajectories with respect to their motion similarity.
- Next, for each trajectory, we compute the velocity by taking differences over consecutive frames in x-y directions.
- In the final step, average magnitude velocities over all trajectories in each cluster are computed frame-wise.

Thus, we get C_j motion clusters per audio source (each source being associated to a different performer's bounding box). Their velocity vectors are resampled in time to match the N STFT frames and arranged in the columns of a matrix $\mathbf{M} \in \mathbb{R}_+^{N \times C}$ where $C = \sum_{j=1}^{J} C_j$.

4.2.2 Model parameter estimation

To illustrate the central idea of the methods given below, assume that the magnitude trajectory of a violinist's bow velocity, given by $\mathbf{m}_{vbow} \in \mathbb{R}_+^N$ is known for a string quartet performance recording, along with the NMF decomposition of the input audio mixtures's magnitude spectrogram, $\mathbf{V} \approx \mathbf{WH}$. We can then try to determine a linear transformation $\boldsymbol{\alpha}_{vbow} \in \mathbb{R}_+^K$ of the activation matrix $\mathbf{H}$ such that $\mathbf{H}^\top \boldsymbol{\alpha}_{vbow}$ is similar to $\mathbf{m}_{vbow}$ with respect to ℓ_2-norm based reconstruction error. This can be formulated as a non-negative least squares (NNLS) cost function. The nonnegativity of $\boldsymbol{\alpha}_{vbow}$ allows for part-based decomposition of $\mathbf{m}_{vbow}$. This can be interpreted as the strength of the contribution of each spectral pattern activation vector, $\mathbf{h}_k$ for reconstruction. We use this later for assigning NMF components to audio sources. Ideally, we expect $\boldsymbol{\alpha}_{vbow}$ to be sparse. In other words, to be concentrated on a few coefficients which indicate that few activations of spectral patterns are linked to bow velocity.

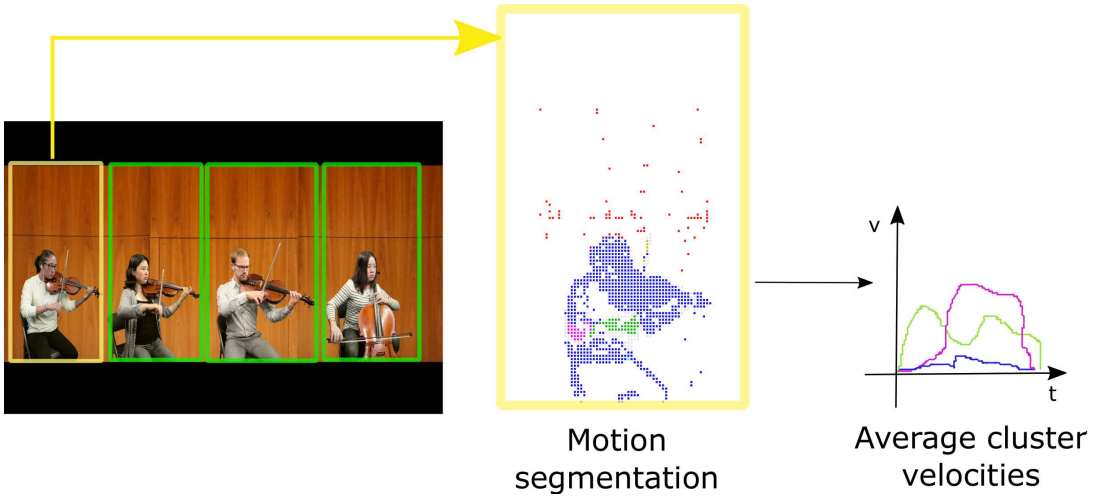


Figure 4.2. Motion Processing Unit: For each bounding box in the video (left), we compute motion segmentation using the multicuts algorithm (Keuper et al., 2015) (centre) and finally, average velocities over each cluster (right). Some clusters in pink, green, blue (foreground) and red (background) are visible. The graph on the right is only a sketch.

Thus, at this step of the algorithm, we determine this linear transformation, denoted by $\boldsymbol{\alpha}_c$, for each velocity vector $\mathbf{m}_c \in \mathbb{R}_+^N$ in $\mathbf{M}$, where $c = 1 \cdots C$, with the expectation that the ones corresponding to the sound-producing motion would be sparse as stated in the illustrative example above.

Defining the nonnegative linear combination coefficient matrix as $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_C]$, the following joint and sequential pathways could be taken for determining $\mathbf{A}$ while minimizing the Frobenius reconstruction error between $\mathbf{M}$ and $\mathbf{H}^\top \mathbf{A}$:

Sequential estimation. Two alternative schemes are considered here:

1. **NMF + NNLS**: After obtaining a blind NMF decomposition of the audio mixture, we perform NNLS where the objective is to determine $\mathbf{A}$ that best reconstructs $\mathbf{M}$ from the given audio activations $\mathbf{H}$. This can be written mathematically as:

$$\begin{aligned} &\text{minimize} && \|\mathbf{M} - \mathbf{H}^\top \mathbf{A}\|_F^2 \\ &\text{subject to} && \mathbf{A} \geq 0, \end{aligned} \tag{4.1}$$

where $\mathbf{A} \geq 0$ denotes nonnegative entries of matrix $\mathbf{A}$. The above formulation is equivalent to solving the NMF problem with $\mathbf{H}$ held constant.

2. **NMF + Sparse NNLS**: Within the previous formulation, concentration of $\boldsymbol{\alpha}_c$ on a few coefficients can be achieved by incorporating a sparsity constraint. This can be achieved through an ℓ_1-regularization term as follows:

$$\begin{aligned} &\text{minimize} && \|\mathbf{M} - \mathbf{H}^\top \mathbf{A}\|_F^2 + \mu \|\mathbf{A}\|_1 \\ &\text{subject to} && \mathbf{A} \geq 0, \end{aligned} \tag{4.2}$$

where μ is a positive constant. Equation (4.2) can be looked at as a sparse-NMF formulation where the basis vectors (here $\mathbf{H}^\top$) are held constant.

Joint NMF-Sparse NNLS. Here we propose a *novel* joint formulation where the cost functions for audio factorization and sparse-NNLS are simultaneously minimized:

$$C(\mathbf{W}, \mathbf{H}, \mathbf{A}) = D_{KL}(\mathbf{V}|\mathbf{WH}) + \lambda \|\mathbf{M} - \mathbf{H}^\top \mathbf{A}\|_F^2 + \mu \|\mathbf{A}\|_1, \tag{4.3}$$

where $D_{KL}(.|.)$ is the Kullback-Leibler divergence and λ is a regularization parameter. Note that it is trivial to minimize the cost function in absence of scaling constraints: $C(\gamma\mathbf{W}, \mathbf{H}/\gamma, \mathbf{A}\gamma) < C(\mathbf{W}, \mathbf{H}, \mathbf{A})$. Taking γ close to zero would lead to degenerate solutions. Therefore, we constrain the columns of $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$ to have unit norm (Eggert and Korner, 2004):

$$\begin{aligned} &\text{minimize} && D_{KL}(\mathbf{V}|\mathbf{WH}) + \lambda \|\mathbf{M} - \mathbf{H}^\top \mathbf{A}\|_F^2 + \mu \|\mathbf{A}\|_1 \\ &\text{subject to} && \mathbf{W} \geq 0, \mathbf{H} \geq 0, \mathbf{A} \geq 0 \\ & && \|\mathbf{w}_k\| = 1, \ \forall k \end{aligned} \tag{4.4}$$

Details regarding the update rules for each variable and implementation are summarized in Algorithm 4.1. Specifically, we use multiplicative update heuristics, as explained by equation 2.15, to derive rules for $\mathbf{H}$, $\mathbf{W}$ and $\mathbf{A}$ given on line (8),

(10) and (13) of Algorithm 1 respectively.[1] The update rule for $\mathbf{W}$ is derived as in (Le Roux et al., 2015). Computationally, in each iteration, updating $\mathbf{W}$ costs $\mathcal{O}(FKN)$, whereas $\mathbf{H}$ and $\mathbf{A}$ updates require $\mathcal{O}(FKN + CKN)$ and $\mathcal{O}(CKN)$, respectively. Thus, the global per iteration complexity is $\mathcal{O}(FKN + CKN)$. Since, usually $C < F$ this is comparable to per iteration cost of NMF *i.e.* $\mathcal{O}(FKN)$.

To avoid confusion and clutter in Algorithm 4.1, we use $\mathbf{\Lambda} = \mathbf{WH}$. Product $\odot$, division and exponents denote element-wise operations, $\mathbf{1}$ denotes a matrix with all entries equal to one and size given by context.

4.2.3 Audio spectral pattern assignment and reconstruction

Once we obtain $\mathbf{A}$, which contains $\boldsymbol{\alpha}_c$ for each of the C velocity clusters, the k-th basis vector is assigned to the j^{th} source if $\text{argmax}_c\, \alpha_{kc}$ belongs to the j^{th} source cluster. Once these assignments are made, we perform soft mask reconstruction for each source (equation 2.16) using $\mathbf{W}_j$ and $\mathbf{H}_j$, the submatrices for spectral patterns and their activations assigned to the j^{th} source by the above-mentioned scheme.

4.3 Results and discussion

The performance of the proposed method is evaluated through tests with two distinct multimodal datasets. General implementation details common to all experiments are detailed below. Some separation results and supplementary material is made available on our companion web page.[2] We evaluate with the following techniques (See Section 4.2.2):

- **LS**: NMF + NNLS;
- **spLS**: NMF + Sparse-NNLS with $\mu = 5$;
- **JLS Rand**: Joint NMF-Sparse NNLS with $\mathbf{W}$ and $\mathbf{H}$ initialized randomly, $\lambda = 0.01$ and $\mu = 0.1$;

[1] Parameter updates are derived in appendix A.

[2] https://goo.gl/y7A5az

Algorithm 4.1. Joint NMF-Sparse NNLS .

Input: $\mathbf{V}, \mathbf{M}, K, \lambda \geq 0, \mu \geq 0$

$\mathbf{W}$, $\mathbf{H}$, $\mathbf{A}$ initialized randomly

$\mathbf{H} \leftarrow \text{diag}(\|\mathbf{w}_1\|, \ldots, \|\mathbf{w}_K\|)\mathbf{H}$

$\mathbf{A} \leftarrow \text{diag}(\|\mathbf{w}_1\|^{-1}, \ldots, \|\mathbf{w}_K\|^{-1})\mathbf{A}$

$\mathbf{W} \leftarrow \mathbf{W}\text{diag}(\|\mathbf{w}_1\|^{-1}, \ldots, \|\mathbf{w}_K\|^{-1})$ ▷ Normalize

$\mathbf{\Lambda} = \mathbf{WH}$

repeat

  $\mathbf{H} \leftarrow \mathbf{H} \odot \dfrac{\mathbf{W}^\top (\mathbf{V} \odot \mathbf{\Lambda}^{-1}) + \lambda \mathbf{A}\mathbf{M}^\top}{\mathbf{W}^\top \mathbf{1} + \lambda \mathbf{A}\mathbf{A}^\top \mathbf{H}}$

  $\mathbf{\Lambda} = \mathbf{WH}$

  $\mathbf{W} \leftarrow \mathbf{W} \odot \dfrac{(\mathbf{\Lambda}^{-1} \odot \mathbf{V})\mathbf{H}^\top + \mathbf{W} \odot \left(\mathbf{1}(\mathbf{W} \odot (\mathbf{1}\mathbf{H}^\top)\right)}{\mathbf{1}\mathbf{H}^\top + \mathbf{W} \odot (\mathbf{1}(\mathbf{W} \odot ((\mathbf{\Lambda}^{-1} \odot \mathbf{V})\mathbf{H}^\top)))}$

  $\mathbf{W} \leftarrow \mathbf{W}\text{diag}(\|\mathbf{w}_1\|^{-1}, \ldots, \|\mathbf{w}_K\|^{-1})$

  $\mathbf{\Lambda} = \mathbf{WH}$

  $\mathbf{A} \leftarrow \mathbf{A} \odot \dfrac{\lambda \mathbf{H}\mathbf{M}}{\lambda \mathbf{H}\mathbf{H}^\top \mathbf{A} + \mu}$

until convergence

return $\mathbf{W}, \mathbf{H}, \mathbf{A}$

- **JLS NMF**: Joint NMF-Sparse NNLS with **W** and **H** initialized using the output obtained after applying NMF to the mixture, $\lambda = 0.01$ and $\mu = 0.1$.

General Implementation Details. For all the experiments, audio spectrograms are computed with a Hamming window of size 4096 and 75% overlap. Thus, we have a $2049 \times N$ matrix where N is the number of STFT frames. Code provided by Févotte and Idier (2011) is used for the standard NMF algorithms. *LS* and *spLS* formulations are implemented using publicly available sparse-NMF code (Le Roux et al., 2015), with sparsity set to zero for *LS*.

Evaluation metrics. We report Signal to Distortion Ratio (SDR) expressed in dB (see equation 3.3), computed using the BSS_EVAL Toolbox version 3.0 (Vincent et al., 2006). NMF for each of the methods is run for 200 iterations. For each mixture, all methods are run 5 times with different random initializations and the reconstruction is performed using a soft mask. SDR is averaged over all runs and mixtures of each set.

Table 4.1. MoCap Dataset (Marchini et al., 2014): SDR for different methods averaged over mixtures of each set. Best SDR is displayed in bold.

Methods	**Mixtures**		
	Set 1 (trios)	Set 2 (duos–diff)	Set 3 (duos–same)
LS	1.0	3.8	0.9
spLS	1.0	3.7	0.8
JLS Rand	1.0	3.8	0.7
JLS NMF	**2.1**	4.7	1.1
sMcNMF	1.3	3.8	**2.2**
Mel NMF	1.0	**5.3**	−0.8

4.3.1 Experiments with motion capture data

These experiments are performed with the EEP dataset. Bow velocity descriptor for each source is used as a substitute for the velocities extracted from moving regions. This is meant to validate the proposed factorization schemes with "ideal" motion features before considering the more challenging video scenario as described in Sec 4.2.1. Thus in this simple case, the total number of clusters in $\mathbf{M}$ is equal to the number of sources in each mixture ($C = J$). For all the proposed methods the number of audio components is set to $(15 \times J)$, *e.g.* $K = 30$ for mixtures with 2 sources.

Table 4.2. URMP Video (Li et al., 2016b): SDR for different methods averaged over mixtures of each set. Best SDR is displayed in bold.

Methods	**Mixtures**		
	Duos	Trios	Quartet
LS	7.0	**3.0**	**0.4**
spLS	**7.1**	2.8	**0.4**
JLS Rand	5.8	1.6	−1.9
JLS NMF	6.9	2.8	0.3
Mel NMF	5.5	1.6	−1.1

We compare with audio-only Mel NMF clustering (Spiertz and Gnann, 2009) and our previous approach, Soft Motion Coupled NMF (sMcNMF).

Discussion: From Table 4.1 we see that the joint approach with audio-NMF initialization outperforms all the other methods for the first set and achieves competitive performance for the second. We observe that for Set 2, the JLS NMF approach outperforms the baseline in 4 out of 6 mixtures. When compared with its random version, it seems to converge both faster and better. Moreover, it appears that sparsity for sequential NNLS does not provide significant improvements over LS.

As expected, spLS attempts to concentrate weight on a few coefficients, but these are not very different from those yielded by LS. Also, as we are only interested in maximum values for component assignment, any existing differences are not visible in the reconstruction.

When confronted with sources having similar motion, the performance of the proposed methods is deemed to degrade. In this respect, the EEP dataset is particularly challenging, as we find multiple mixture segments with similar motion. In such cases information such as bow inclination (used by sMcNMF) proves to be quite useful. It is worth mentioning that for some mixtures, all the proposed methods outperform the baselines by a large margin.

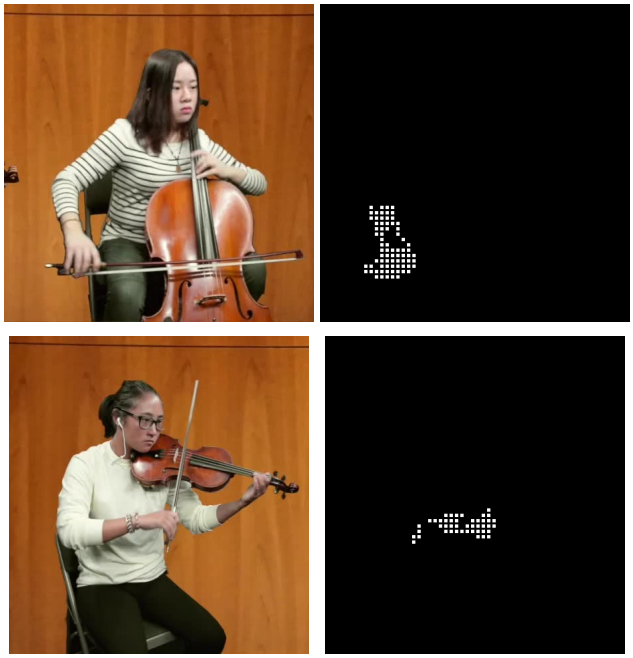
4.3.2 Experiments with videos

In this second series of experiments, we apply the proposed methods to videos. As no standard dataset exists for such a task, we consider the only publicly available example video from the URMP dataset (Li et al., 2016b).[3] We are provided with video recording of a string quartet performance and the separate audio tracks for each player. We consider a 5s excerpt from 30-35s and compute the motion trajectory segmentation for each moving region bounding box (as depicted in Fig. 4.2) using publicly available binary from (Keuper et al., 2015) with default parameter setting. The calculated velocity trajectories are resampled to match N STFT frames. We consider all two, three and four source combinations, denoted by **Duos** (6 mixtures), **Trios** (4 mixtures) and **Quartet** (1 mixture) in Table 4.2. We compare only with Mel-NMF as sMcNMF or the other NMF-based methods are not designed to deal with generic videos. As in the previous case, K is set to $(15 \times J)$ for all methods.

Discussion: The efficacy of the proposed methods is seen from the results in Table 4.2, with particular emphasis on good initialization for the joint approach. The

[3] The full dataset is yet to be released. Sample video 32- The Art of the Fugue can be found at http://www.ece.rochester.edu/projects/air/projects/datasetproject.html

Figure 4.3. Localization results for the first frame of cello and violin. The clusters corresponding to the hand *i.e.* the bowing motion (in white) have been identified in both cases.



quartet is a particularly difficult case where we observed that none of the methods work consistently well over all five runs and the SDR variance is high.

Unlike the previous experiment, here we have multiple velocity clusters to choose from, which makes the problem considerably more difficult. We note that the methods deal reasonably well with low velocity unrelated/noisy clusters as they are not strongly related to any audio activation. Interestingly, it is possible to identify motion trajectory clusters responsible for the sound of each source using $\mathbf{A}$. As indicated by $\text{argmax}_c\, \alpha_{kc}$, for each source we can determine the velocity cluster with maximum audio component assignments. These localization results are illustrated in Fig. 4.3 for a mixture of violin and cello. This provides additional evidence for our hypothesis and the proposed estimation methods.

4.4 Conclusion

To summarize, we have proposed novel motion-assisted methods for audio source separation. This was done by exploiting features encoding physical excitation information in both modalities. We aim to determine non-negative coefficients to regress motion from audio activations. In addition to demonstrating the usefulness of this idea through sequential techniques, we present and derive algorithm for an original joint formulation. While the extension to audio denoising is straightforward, in their current form, the methods cannot deal with high amplitude noisy visual motion.

For the particular case of musical mixtures, score information would prove to be very beneficial in guiding source separation. It can certainly be incorporated within the present framework, which will be a topic for further study. Several other loss functions and non-linear methods for establishing similarity could be experimented with for better performance and wider applicability.

We have shown the usefulness of motion–audio correlation in the context of audio source separation. Its contribution to robust source disambiguation for acoustically similar objects is worth highlighting. These developments complement the use of other types of auxiliary information such as score (Fritsch and Plumbley, 2013), audio timbre (Spiertz and Gnann, 2009), text (Le Magoarou et al., 2015), humming (Smaragdis and Mysore, 2009), *etc.* A source of side information that is particularly relevant to this thesis and readily available to visual systems is object appearance. In the chapters that follow, we aim to develop methods that use audio–appearance co-occurrence to tackle various AV scene understanding tasks.

5

Weakly supervised representation learning for AV events

Synopsis

Audio-visual representation learning is an important task from the perspective of designing machines with the ability to understand complex events. To this end, we propose a novel multimodal framework that instantiates multiple instance learning. We show that the learnt representations are useful for classifying events and localizing their characteristic audio-visual elements. The system is trained using only video-level event labels without any timing information. An important feature of our method is its capacity to learn from unsynchronized audio-visual events. We achieve state-of-the-art results on a large-scale dataset of weakly-labeled audio event videos. Visualizations of localized visual regions and audio segments substantiate our system's efficacy, especially when dealing with noisy situations where modality-specific cues appear asynchronously.

5.1 Introduction

In this chapter, we propose a framework to learn AV representations from large unconstrained video collections involving generic AV objects. As we show later, such representations are particularly useful for identifying and characterizing events that

can be perceived via distinct audio and visual cues, for *e.g.* a ringing phone or a car passing by. This complements work discussed in earlier chapters wherein we exploited motion-audio correlation for tackling visually-indicated sounds. While such correlations prove to be effective for audio source disambiguation and visual motion localization, they do not suffice for tackling problems such as object classification and localization. For instance, two music instruments such as a guitar and mandolin that are visually and acoustically very different may be "struck" in a similar fashion. Consequently, motion–audio relations cannot be used for detecting these objects in either of the modalities. Moreover, they may not always exist or be visible in "in-the-wild" videos. This is particularly true for objects such as car, phone *etc.*, thus requiring a complementary multimodal association mechanism, namely appearance–audio co-occurrence. We leverage this in large datasets to build a more comprehensive AV analysis system capable of performing classification, localization and separation.

Obtaining precisely annotated data for performing the forgoing tasks is an expensive endeavor, made even more challenging by multimodal considerations. The annotation process for object localization in both the modalities is not only error prone and time consuming but also subjective to an extent. Often, event boundaries in audio, extent of video objects or even their presence is ambiguous. Thus, we opt for a weakly-supervised learning approach using data with only global video-level event labels, that is labels given for whole video documents without timing information.

Problem description. To motivate our tasks and method, consider a video labeled as "train horn", depicted in Fig. 5.1. Assuming that the train is both visible and audible at some time in the video, in addition to identifying the event, we are interested in learning representations that help us answer the following[1]:

- *Where is the visual object or context that distinguishes the event?* In this case it might be the train (object) or tracks, platform (context) *etc.* We are thus aiming for their spatio-temporal localization in the image sequence (this chapter).

[1] The parentheses at the end of each bullet point indicate the chapter in which we tackle the corresponding question

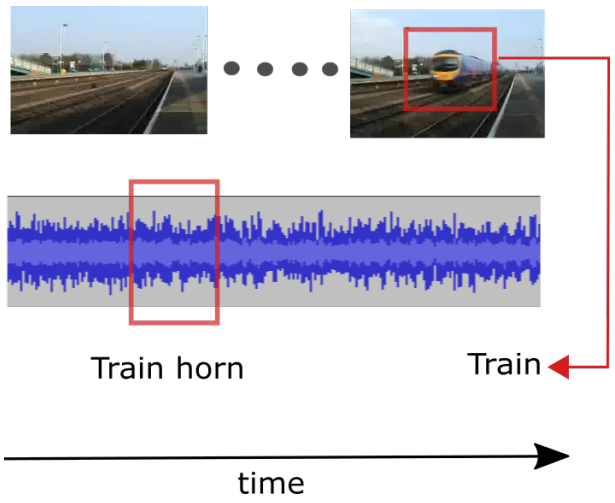


Figure 5.1. Pictorial representation of the problem: Given a video labeled as "train horn", we would like to: (i) identify the event, (ii) localize both, its visual presence and the temporal segment(s) containing the characteristic sound, and (iii) segregate the characteristic audio cue from the background. Note that the train horn may sound before the train is visible. Our model can deal with such unsynchronized AV events.

- *When does the sound event occur?* Here it is the train horn. We thus want to temporally localize the audio event (this chapter).
- *Where is the audio object?* Here we are interested in audio source extraction *i.e.* segregating the source of interest from the background sounds (next chapter).

The variety of noisy situations that one may encounter in unconstrained environments or videos adds to the difficulty of this very challenging problem. Apart from modality-specific noise such as visual clutter, lighting variations and low audio signal-to-noise ratio, in real-world scenarios the appearance of audio and visual elements characterizing the event are often unsynchronized in time. This is to say that the train horn may sound before or after the train is visible, as in previous example. In the extreme, not so rare case, the train may not appear at all. The latter is also commonly referred to as "off–screen" audio. We are interested in designing a system to tackle the aforementioned questions and situations.

Prior research has utilized audio and visual modalities for classification and localiz-

ation tasks in various contexts. Fusing modality-specific hand-crafted or deep features has been a popular approach for problems such as multimedia event detection and video concept classification (Jiang et al., 2009b; Chang et al., 2007; Jiang et al., 2018, 2013). On the other hand, AV correlations have been utilized for localization and representation learning in general, through feature space transformation techniques such as canonical correlation analysis (CCA) (Izadinia et al., 2013; Kidron et al., 2005) or deep networks (Owens et al., 2016a,b; Arandjelović and Zisserman, 2017a,b; Andrew et al., 2013). However, a unified multimodal framework for our task, that is learning data representations for simultaneously identifying real world events and the AV cues depicting them has not been extensively studied in previous works.

Chapter outline. We begin by briefly mentioning connections and distinctions with related works in Section 5.2. This is followed by a description of the proposed framework and its instantiations for tackling classification and localization in Section 5.3. Finally, we validate the usefulness of the learnt representations for these tasks with a thorough analysis in Section 5.4.

5.2 Related work

To position our work, we briefly discuss some relevant literature that employs weakly supervised learning for object localization and event detection in computer vision and machine listening, respectively. We also dileneate several distinctions between the present study and recent multimodal deep learning approaches.

5.2.1 Visual object localization and classification

There is a long history of works in computer vision applying weakly supervised learning for object localization and classification. MIL techniques have been extensively used for this purpose (Zhang et al., 2006; Bilen et al., 2014b; Oquab et al., 2015; Kantorov et al., 2016; Bilen and Vedaldi, 2016; Zhou et al., 2016; Cinbis et al., 2017). Typically, each image is represented as a set of regions. Positive images contain at least one

region from the reference class while negative images contain none. Latent structured output methods, *e.g.*, based on support vector machines (SVMs) (Bilen et al., 2014a) or conditional random fields (CRFs) (Deselaers et al., 2010), address this problem by alternating between object appearance model estimation and region selection. Some works have focused on better initialization and regularization strategies (Kumar et al., 2010; Song et al., 2014; Cinbis et al., 2017) for solving this non-convex optimization problem.

Owing to the exceptional success of convolutional neural networks (CNNs) in computer vision, recently, several approaches have looked to build upon CNN architectures for embedding MIL strategies. These include the introduction of operations such as max pooling over regions (Oquab et al., 2015), global average pooling (Zhou et al., 2016) and their soft versions (Kolesnikov and Lampert, 2016). Another line of research consists in CNN-based localization over class-agnostic region proposals (Gkioxari et al., 2015; Bilen and Vedaldi, 2016; Kantorov et al., 2016) extracted using a state-of-the-art proposal generation algorithm such as EdgeBoxes (Zitnick and Dollár, 2014), Selective Search (Uijlings et al., 2013), *etc.* These approaches are supported by the ability to extract fixed size feature maps from CNNs using region-of-interest (Girshick, 2015) or spatial pyramid pooling (He et al., 2015). Our work is related to such techniques. We build upon ideas from the two-stream architecture (Bilen and Vedaldi, 2016) for classification and localization.

State-of-the-art end-to-end object detection networks such as Faster RCNN (Ren et al., 2015) and its instance segmentation extension Mask RCNN (He et al., 2017) incorporate proposal generation as part of the system (region proposal network) instead of a separate stage. Nonetheless, these approaches require label annotations for different regions. It is also worth mentioning that some works have extended class-agnostic proposal generation from 2D images to video tube proposals for tasks such as action localization (Van Gemert et al., 2015) and object detection (Oneata et al., 2014). However, these involve a computationally expensive pipeline preventing large-scale usage.

5.2.2 Audio event detection

A significant amount of literature exists on supervised audio event detection (AED) (Mesaros et al., 2015; Zhuang et al., 2010; Adavanne et al., 2017; Bisot et al., 2017). However, progress with weakly labeled data in the audio domain has been relatively recent. An early work (Kumar and Raj, 2016) showed the usefulness of MIL techniques to audio using SVM and neural networks.

The introduction of the weakly-labeled audio event detection task in the 2017 DCASE challenge (Mesaros et al., 2017)[2], a challenge on Detection and Classification of Acoustic Scenes and Events (DCASE), along with the release of Google's AudioSet data[3] (Gemmeke et al., 2017), has led to accelerated progress in the recent past. AudioSet is a large-scale weakly-labeled dataset of audio events collected from YouTube videos. A subset of this data was used for the DCASE 2017 task on large-scale AED for smart cars.[4] Several submissions to the task utilized sophisticated deep architectures with attention units (Xu et al., 2017), as well as max and softmax operations (Salamon et al., 2017). Another recent study introduced a CNN with global segment-level pooling for dealing with weak labels (Kumar et al., 2017). While we share with these works the high-level goal of weakly-supervised learning, apart from our multimodal design, our audio sub-module, as discussed in the next section, is significantly different.

5.2.3 Differences with recent AV deep learning studies

We formulate the problem as a MIL task using class-agnostic proposals from both video frames and audio. This allows us to simultaneously solve the classification and localization problems. Finally, by construction, our framework deals with the difficult case of asynchronous AV events. This is significantly different from recent multimodal deep learning based studies on several counts: Contrary to prior works, where unsupervised representations are learnt through audio–image correlations (tem-

[2] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/

[3] https://research.google.com/audioset/

[4] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-large-scale-sound-event-detection

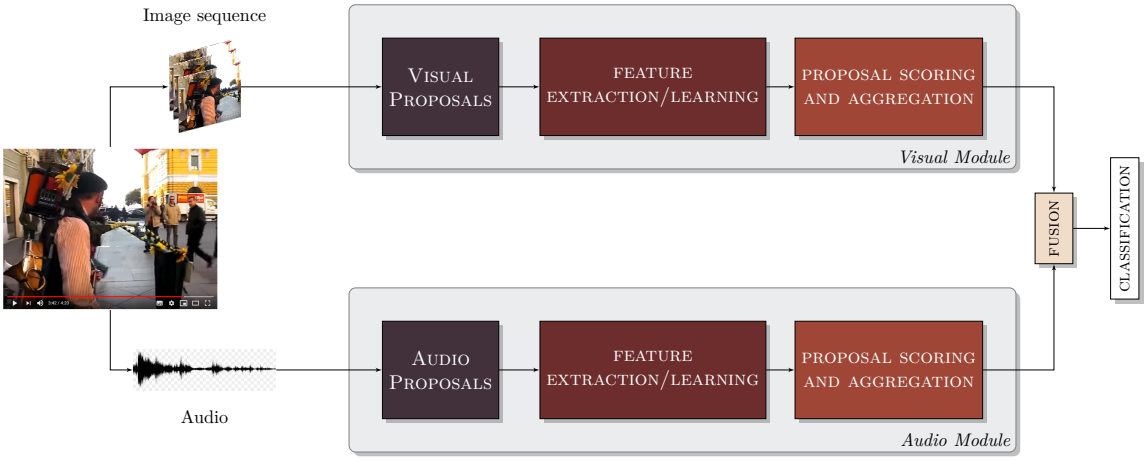


Figure 5.2. High level view of the proposed approach: Given a video captured using a single microphone and camera, we propose the depicted framework for weakly supervised representation learning.

poral co-occurrence), we adopt a weakly-supervised learning approach using event classes. Unlike (Arandjelović and Zisserman, 2017b; Owens et al., 2016a,b), we focus on localizing both, discriminative audio and visual components for real-world events.

5.3 Proposed framework and its instantiation

The tasks under consideration can be naturally interpreted as MIL problems (Dietterich et al., 1997b). 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. In our case, a video can be seen as a labeled bag, containing a collection of visual and audio proposals. The term *proposal* refers to image or audio "parts" that may potentially constitute the object of interest. This step is at the core of our approach.

The key idea, as illustrated in Fig. 5.2, is to extract features from generated proposals and transform them for: (1) scoring each according to their relevance for class labels; (2) aggregating these scores in each modality and fusing them for video-level classification. This not only allows us to train both the sub-modules together

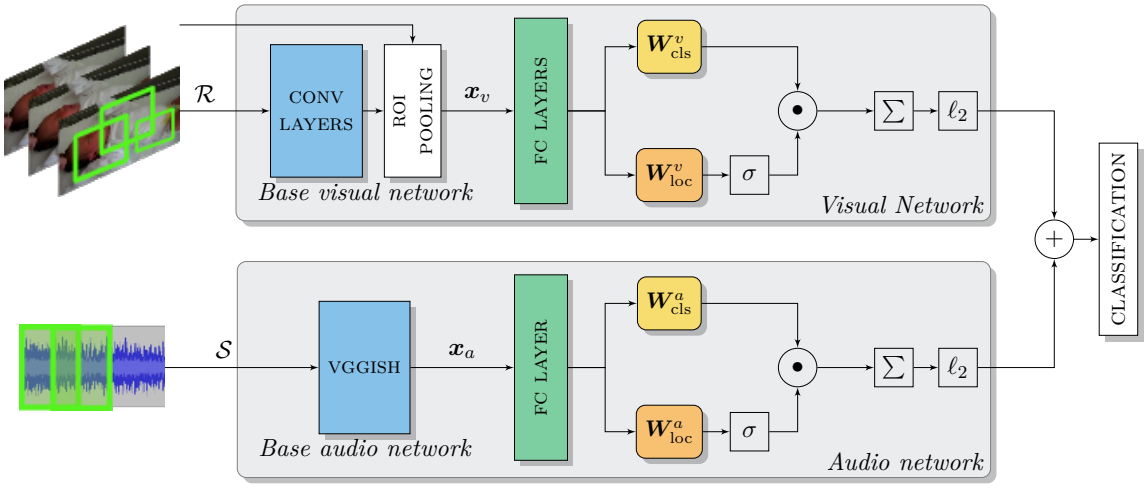


Figure 5.3. Module design: Given a video, we consider the depicted pipeline for going from audio and visual proposals to localization and classification. Here $\boldsymbol{W}_{\text{cls}}$ and $\boldsymbol{W}_{\text{loc}}$ refer to the fully-connected classification and localization streams respectively; σ denotes softmax operation over proposals for each class, $\odot$ refers to element-wise multiplication; Σ to a summation over proposals and ℓ_2 to a normalization of scores. During training we freeze the weights of blocks denoted in blue.

through weak-supervision but also enables localization using the proposal relevance scores. Moreover, use of both the modalities with appropriate proposals makes the system robust against noisy scenarios. In the current and following chapters, we present different task-specific variants of this general framework.

We now formalize the design of each building block to specifically tackle event classification, visual object and audio event localization. An overview is provided in Fig. 5.3. We model a video V as a bag of M selected image regions, $\mathcal{R} = \{r_1, r_2, \ldots, r_M\}$, obtained from sub-sampled frames and T audio segments, $\mathcal{S} = \{s_1, s_2, \ldots, s_T\}$. Given N such training examples, $\mathcal{V} = \{V^{(n)}\}_{n=1}^{N}$, organized into C classes, our goal is to learn a representation to jointly classify and localize image regions and audio segments that characterize a class. Each block from proposal generation to classification is discussed below in detail.

5.3.1 Generating proposals and extracting features

Visual Proposals. Generating proposals for object containing regions from images is at the heart of various visual object detection algorithms (Wang et al., 2013; Girshick et al., 2014). As our goal is to spatially and temporally localize the most discriminative region pertaining to a class, we choose to apply this technique over sub-sampled video frame sequences. In particular, we sub-sample the extracted frame sequences of each video at a rate of 1 frame per second. This is followed by class-agnostic region proposal generation on the selected frames using EdgeBoxes (Zitnick and Dollár, 2014). This proposal generation method builds upon the insight that the number of contours entirely inside a box is indicative of the likelihood of an object's presence. Its use in our pipeline is motivated by experiments confirming better performance in terms of speed/accuracy tradeoffs over most competing techniques (Hosang et al., 2014). EdgeBoxes additionally generates a confidence score for each bounding box which reflects the box's "objectness". To reduce the computational load and redundancy, we use this score to select the top M_{img} proposals from each sampled image, I, and use them for feature extraction. Hence, given a 10 second video, the aforementioned procedure would leave us with 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 V, using a convolutional neural network altered with a region-of-interest (RoI) pooling layer. An RoI layer works by computing fixed size feature maps (*e.g.* 6×6 for `caffenet` (Krizhevsky et al., 2012)) from regions of an image using max-pooling (Girshick, 2015). This helps to ensure compatibility between convolutional and fully connected layers of a network when using regions of varying sizes. Moreover, unlike Region-based CNN (RCNN) (Girshick et al., 2014), the shared computation for different regions of the same image using Fast-RCNN implementation (Girshick, 2015) leads to faster processing. In Fig. 5.3 we refer to this feature extractor as the base visual network. In practice, feature vectors $\boldsymbol{x}_v(\cdot)$ are extracted after RoI pooling layer and passed through two fully connected layers, which are fine-tuned during training. Typically, standard CNN architectures pre-trained on ImageNet

(Deng et al., 2009) classification are used for the purpose of initializing network weights.

Audio Temporal Segment Proposals. We first represent the raw audio waveform as a log-Mel spectrogram (Davis and Mermelstein, 1990). Each proposal is then obtained by sliding a fixed-length window over the obtained spectrogram along the temporal axis. The dimensions of this window are chosen to be compatible with the audio feature extractor. For our system we set the proposal window length to 960ms and stride to 480ms.

We use a VGG-style deep network known as `vggish` for base audio feature extraction. Inspired by the success of CNNs in visual object recognition Hershey et al. (2017) introduced this state-of-the-art audio feature extractor as an audio parallel to networks pre-trained on ImageNet for classification. `vggish` has been pre-trained on a preliminary version of YouTube-8M (Abu-El-Haija et al., 2016) for audio classification based on video tags. It stacks 4 convolutional and 2 fully connected layers to generate a 128 dimensional embedding, $\boldsymbol{x}_a(s_t; V) \in \mathbb{R}^{128}$ for each input log-Mel spectrogram segment $s_t \in \mathbb{R}^{96\times64}$ with 64 Mel-bands and 96 temporal frames. Prior to proposal scoring, the generated embedding is passed through a fully-connected layer that is learnt from scratch.

5.3.2 Proposal scoring network and fusion

So far, we have extracted base features for each proposal in both the modalities and passed them through fully connected layers in their respective modules. Equipped with this transformed representation of each proposal, we use the two-stream architecture proposed by Bilen and Vedaldi (2016) for scoring each of them with respect to the classes. There is one scoring network of the same architecture for each modality as depicted in Fig. 5.3. Thus, for notational convenience, we generically denote the set of audio or visual proposals for each video by $\mathcal{P}$ and let proposal representations before the scoring network be stacked in a matrix $Z \in \mathbb{R}^{|\mathcal{P}|\times d}$, where d denotes the dimensionality of the audio/visual proposal representation.

The architecture of this module consists of parallel classification and localiza-

tion streams. The former classifies each region by passing Z through a linear fully connected layer with weights W_{cls}, giving a matrix $A \in \mathbb{R}^{|\mathcal{P}|\times C}$. On the other hand, the localization layer passes the same input through another fully-connected layer with weights W_{loc}. This is followed by a softmax operation over the resulting matrix $B \in \mathbb{R}^{|\mathcal{P}|\times C}$ in the localization stream. The softmax operation on each element of B can be written as:

$$[\sigma(B)]_{pc} = \frac{e^{b_{pc}}}{\sum_{p'=1}^{|\mathcal{P}|} e^{b_{p'c}}}, \ \forall (p,c) \in (1,|\mathcal{P}|) \times (1,C). \tag{5.1}$$

This allows the localization layer to choose the most relevant proposals for each class. Subsequently, the classification stream output is weighted by $\sigma(B)$ through element-wise multiplication: $D = A \odot \sigma(B)$. Class scores over the video are obtained by summing the resulting weighted scores in D over proposals. Note that the dataset we use, by construction, allows a region or segment to belong to multiple classes. Hence, we do not opt for softmax on the classification stream, as done in (Bilen and Vedaldi, 2016).

After performing the above stated operations for both audio and visual sub-modules, in the final step, the global video-level scores are ℓ_2 normalized and added. In preliminary experiments we found this to work better than addition of unnormalized scores. We hypothesize that the system trains better because ℓ_2 normalization ensures that the scores being added are in the same range.

5.3.3 Classification loss and network 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 $\mathcal{R}^{(n)}$ and audio segments $\mathcal{S}^{(n)}$. After performing the described operations on each modality separately, the ℓ_2 normalized scores are added and represented by $\varphi(V^{(n)}; \boldsymbol{w}) \in \mathbb{R}^C$, with all network weights and biases denoted by $\boldsymbol{w}$. All the weights, including and following the fully-connected layer processing stage

for both the modalities, are included in $\boldsymbol{w}$. Note that both sub-modules are trained jointly. The network is trained using the multi-label hinge loss on a batch of size B:

$$L(\boldsymbol{w}) = \frac{1}{CB} \sum_{n=1}^{B} \sum_{c=1}^{C} \max\left(0, 1 - y_c^{(n)} \varphi_c(V^{(n)}; \boldsymbol{w})\right). \tag{5.2}$$

5.4 Experimental validation

Dataset. We use the recently introduced dataset for the DCASE challenge on large-scale weakly supervised sound event detection for smart cars (Mesaros et al., 2017). This is a subset of Audioset (Gemmeke et al., 2017) which contains a collection of weakly-annotated unconstrained YouTube videos of vehicle and warning sounds spread over 17 classes. It is categorized as follows:

- *Warning sounds*: Train horn, Air horn, Truck horn, Car alarm, Reversing beeps, Ambulance (siren), Police car (siren), Fire engine/fire truck (siren), Civil defense siren, Screaming.
- *Vehicle sounds*: Bicycle, Skateboard, Car, Car passing by, Bus, Truck, Motorcycle, Train.

This multi-label dataset contains 51,172 training, 488 validation and 1103 testing samples. Despite our best efforts, due to YouTube and video downloader issues, some videos were unavailable, not downloadable or contained no audio. This left us with 48,719 training, 462 validation and 1030 testing clips. It is worth mentioning that the training data is highly unbalanced with the number of samples for the classes ranging from 175 to 24K. We provide details regarding different classes and the number of samples in each in appendix B. To mitigate the negative effect of this imbalance on training, we introduce some balance by ensuring that each training batch contains at least one sample from some or all of the under-represented classes. Briefly, each batch is generated by first randomly sampling labels from a specific list, followed by fetching examples corresponding to the number of times each label is sampled.

This list is generated by ensuring higher but limited presence of classes with more examples. We use a publicly available implementation for this purpose (Xu et al., 2017).[5]

Baselines. To our best knowledge, there is no prior work on deep architectures that perform the task of weakly supervised classification and localization for unsynchronized AV events. Our task and method are substantially different from recently proposed networks like L3 (Arandjelović and Zisserman, 2017a,b) which are trained using synchronous AV pairs on a large collection of videos in a self-supervised manner. However, we designed several strong baselines for comparison and an ablation study. In particular, we compare against the following networks:

1. AV One-Stream Architecture: Applying MIL in a straight-forward manner, we could proceed only with a single stream. That is, we can use the classification stream followed by a max operation for selecting the highest scoring regions and segments for obtaining global video-level scores. As done in (Bilen and Vedaldi, 2016), we choose to implement this as a multimodal MIL-based baseline. We replace the *max* operation by the *log-sum-exponential* operator, its soft approximation. This has been shown to yield better results (Bilen et al., 2014b). The scores on both the streams are ℓ_2 normalized before addition for classification. This essentially amounts to removing from Fig. 5.3 the localization branches and replacing the summation over proposals with the soft-maximum operation described above. To avoid any confusion, please note that we use the term 'stream' to refer to classification and localization parts of the scoring network.

2. Visual-Only and Audio-Only Networks: These networks only utilize one of the modalities for classification. However, note that there are still two streams for classification and localization, respectively. For a fair comparison and ablation study we train these networks with ℓ_2 normalization. In addition, for completeness we also implement Bilen *et al.*'s architecture for weakly supervised deep detection networks (WSDDN) with an additional softmax on the classification stream. As

[5] https://github.com/yongxuUSTC/dcase2017_task4_cvssp/blob/master/data_generator.py

the scores are in the range [0,1], we train this particular network with C binary log-loss terms (Bilen and Vedaldi, 2016). When discussing results we refer to this system as WSDDN-Type.

3. CVSSP Audio-Only (Xu et al., 2017): This state-of-the-art method is the DCASE 2017 challenge winner for the audio event classification sub-task. The system is based on Gated convolutional RNN (CRNN) for better temporal modeling and attention-based localization. They use no external data and training/evaluation is carried out on all the samples. We present results for both their winning fusion system, which combines prediction of various models and Gated-RCNN model trained with log-Mel spectrum.

Implementation details. All systems except that of (Xu et al., 2017), including variants, are implemented in Tensorflow. They were trained for 25K iterations using Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-5} and a batch size of 24. We use the MATLAB implementation of EdgeBoxes for generating region proposals, obtaining approximately 100 regions per video with $M_{\text{img}} = 10$ and a duration of 10s. The implementation is used with default parameter setting. Base visual features, $\boldsymbol{x}_v \in \mathbb{R}^{9216}$ are extracted using `caffenet` (Krizhevsky et al., 2012) with pre-trained ImageNet weights and RoI pooling layer modification (Girshick, 2015). With 6×6 RoI pooling we get a 9216 ($= 256 \times 6 \times 6$) dimensional feature vector. For this, the Fast-RCNN Caffe implementation is used (Girshick, 2015). The fully connected layers, namely fc_6 and fc_7, each with 4096 neurons, are fine-tuned, with 50% dropout during training.

For audio, each recording is resampled to 16 kHz before processing. Log-Mel spectrum over the whole file is computed with a window size of 25ms and 10ms hop length. The resulting spectrum is chunked into segment proposals using a 960–ms window with a 480–ms stride. For a 10–second recording, this yields 20 segments of size 96×64. We use the official Tensorflow implementation of `vggish`.[6] The base audio features extracted from `vggish` are run through a fully connected layer with

[6] https://github.com/tensorflow/models/tree/master/research/audioset

128 neurons. This layer is learnt from scratch along with the scoring networks during training.

Metrics. The baselines and proposed systems are evaluated on the micro-averaged F1 score. The term micro-averaging implies that the F1 score is computed using a global count of total true positives (TP), false negatives (FN) and false positives (FP). This was the official metric used by DCASE 2017 smart cars task for ranking systems (Mesaros et al., 2017). It is mathematically defined as:

$$\text{F1} = \frac{2P \times R}{P + R}, \text{ where } P = \frac{\text{TP}}{\text{TP} + \text{FP}}, \; R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.3}$$

Here, P and R are the precision and recall measures, respectively. The score thresholds for each system are determined by tuning over validation data to maximize F1 score for each class. They are then applied to the test data for final predictions. For further insight, we also report here the F1 scores for each class.

To quantify sound event detection performance, in addition to the segment-wise aggregated F1 score, the error rate (ER) is computed. This measure compares the ground truth and estimated output using one second long sub–segments. For each sample, ER is computed as:

$$ER = \frac{S + D + I}{N_{\text{e}}}, \tag{5.4}$$

where N_{e} denotes total number of reference events in all segments and S, D, I correspond to substitution, deletion and insertion errors, respectively. For more details, please see (Mesaros et al., 2016).[7]

Results and discussion

Quantitative results. We show in Table 5.1 the micro-averaged F1 scores for each of the systems described in this chapter. In particular, system (a) in Table 5.1 is the proposed AV system and (b)-(e) present its variants which are also treated

[7] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/metrics

as baselines, (f)-(g) denote results from CVSSP team (Xu et al., 2017), winners of the DCASE AED for smart cars audio event tagging task. We outperform all the approaches by a significant margin. Among the multimodal systems, the two-stream architecture performs much better than the one-stream counter-part, designed with only a classification stream and soft-maximum for region selection. On the other hand, the state-of-the-art CVSSP fusion system, which combines predictions of various models, achieves a better precision than the other methods. Several important and interesting observations can be made by looking at these results in conjunction with the class-wise scores reported in Table 5.2.

Most importantly, the results emphasize the complementary role of visual and audio sub-modules for this task. To see this, we could categorize the data into two sets: (i) classes with clearly defined AV elements, for instance car, train, motorcycle; (ii) some warning sounds such as, *e.g.*, reverse beeping, screaming, air horn, where the visual object's presence is ambiguous. The class-wise results of the video only system are a clear indication of this split. Well-defined visual cues enhance the performance of the proposed multimodal system over audio-only approaches, as video frames carry vital information about the object. On the other hand, in the case of warning sounds, video frames alone are insufficient as evidenced by results for the video-only system. In this case, the presence of audio assists the system in arriving at the correct prediction. The expected AV complementarity is clearly established through these results.

Note that for some warning sounds the CVSSP method achieves better results. In this regard, we believe better temporal modeling for our audio system could lead to further improvements. In fact, we currently operate with a coarse temporal window of 960ms, which might not be ideal for all audio events. RNNs could also be used for further improvements. We think such improvements are orthogonal and were not the focus of this study. We also observe that results for under-represented classes in the training data such as air horn and reversing beeps are relatively lower. This can possibly be mitigated through data augmentation strategies.

Finally, we show the sound event detection performance in Table 5.3. The results are computed by simply thresholding the two–stream output from the audio sub–

Table 5.1. Results on DCASE smart cars task test set. We report here the micro-averaged F1 score, precision and recall values and compare with state-of-the-art. TS is an acronym for two-stream.

	System	F1	Precision	Recall
(a)	Proposed AV Two Stream	**64.2**	59.7	**69.4**
(b)	TS Audio-Only	57.3	53.2	62.0
(c)	TS Video-Only	47.3	48.5	46.1
(d)	TS Video-Only WSDDN-Type (Bilen and Vedaldi, 2016)	48.8	47.6	50.1
(e)	AV One Stream	55.3	50.4	61.2
(f)	CVSSP - Fusion system (Xu et al., 2017)	55.6	**61.4**	50.8
(g)	CVSSP - Gated-CRNN-logMel (Xu et al., 2017)	54.2	58.9	50.2

module at $\tau = 0$ for the predicted label(s). We note that the results are comparable with the best performing CVSSP system. Note that the winning system for this subtask from Lee et al. (2017) employs an ensemble method to optimally weigh multiple learned models, using ER as the performance metric to make the final selection. No such fine tuning is performed in our case.

Qualitative results. Fig. 5.5 displays several video frames from different evaluation videos where we achieve good visual localization for various objects. The heatmaps shown below the images denote image region (top) and audio segment detection scores (bottom) for the reference class in sub-figure captions. The x-axis for the former denotes all the proposals from the subsampled images (in temporal order), whereas for audio it denotes the overlapping segment time-stamps. The display uses 'hot' colormap where black is 0 and white is 1, as depicted in Fig. 5.7. We see that the most discriminative proposals are found at different time instants for each modality. Some more positive examples and typical failure modes are shown in Fig. 5.6.

A by-product of our design is the ability to deal with asynchronous AV events. We present in Fig. 5.7 two examples to demonstrate this. In the first case **A**, the

Table 5.2. Class-wise comparison on test set using F1 scores. We use TS, OS and FS as acronyms to refer to two-stream, one-stream and fusion system, respectively.

System	Vehicle Sounds								Warning Sounds								
	bik	bus	car	car-pby	mbik	skt	trn	trk	air-hrn	amb	car-alm	civ-def	f-eng	pol-car	rv-bps	scrm	trn-hrn
Proposed AV TS	**75.7**	**54.9**	**75.0**	**34.6**	**76.2**	**78.6**	82.0	**61.5**	40.0	**64.7**	53.9	80.4	64.4	49.2	36.6	81.1	47.1
TS Audio-Only	42.1	38.8	69.8	29.6	68.9	64.9	78.5	44.0	40.4	58.2	53.0	79.6	61.0	51.4	42.9	72.1	46.9
TS Video-Only	72.5	52.0	61.2	15.0	54.1	64.2	73.3	49.7	12.0	33.9	13.5	68.6	46.5	19.8	21.8	44.1	32.1
AV OS	68.2	53.6	74.1	25.6	67.1	74.4	**82.8**	52.8	28.0	54.7	20.6	76.6	60.4	56.3	18.8	49.4	36.2
CVSSP - FS	40.5	39.7	72.9	27.1	63.5	74.5	79.2	52.3	**63.7**	35.6	**72.9**	**86.4**	**65.7**	**63.8**	**60.3**	**91.2**	**73.6**

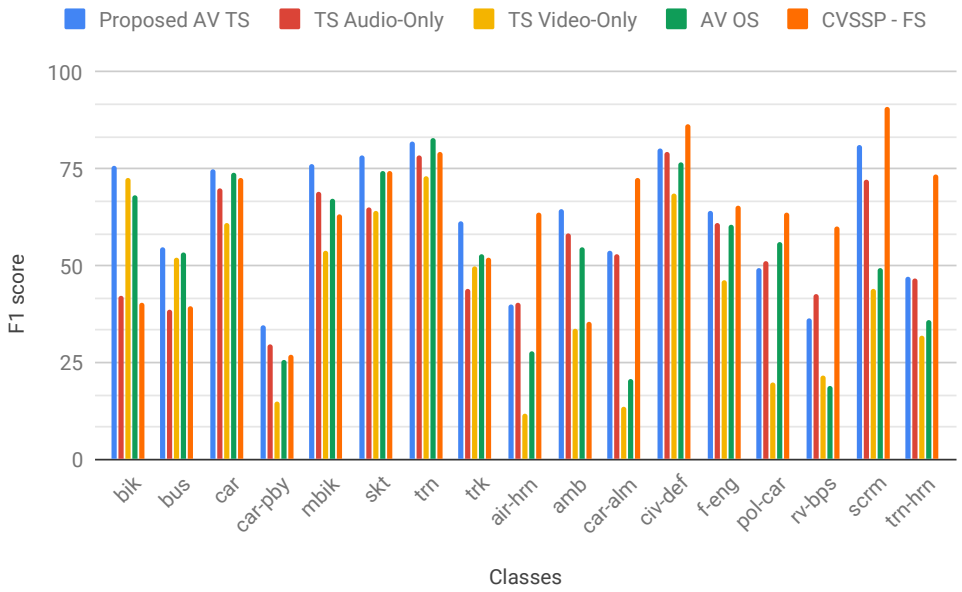


Figure 5.4. Bar graph illustration of class-wise F1 scores from Table 5.2

Table 5.3. F1 score and error rate for sound event detection task

	System	F1	ER
(a)	Proposed AV Two Stream	51.0	0.76
(b)	TS Audio-Only	48.5	0.78
(c)	CVSSP - Fusion system (Xu et al., 2017)	51.8	0.73
(d)	CVSSP - Gated-CRNN-logMel (Xu et al., 2017)	47.5	0.78
(e)	SNU - Ensemble method (Lee et al., 2017)	**55.5**	**0.66**

sound of a car's engine is heard in the first two seconds followed by music. The normalized audio localization heatmap at the bottom displays the scores assigned to each temporal audio segment, s_t by the car classifier. The video frames placed above are roughly aligned with the audio temporal axis to show the video frame at

the instant when the car sounds and the point where the visual network localizes. The localization is displayed through a yellow bounding box. To better understand the system's output, we modulate the opacity of the bounding box according to the system's score for it. Higher the score, more visible the bounding box. As expected, we do not observe any yellow edges in the first frame. Clearly, there exists temporal asynchrony, where the system locks onto the car, much later, when it is completely visible. **B** depicts an example, where due to extreme lighting conditions the visual object is not visible. Here too, we localize the audio object and correctly predict the 'motorcycle' class.[8]

5.5 Conclusion

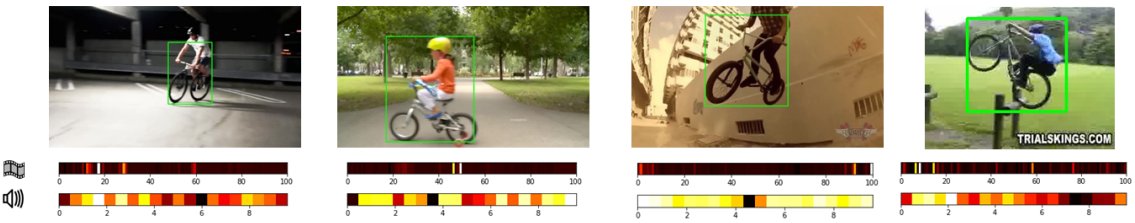
We have proposed a novel approach based on a deep multimodal architecture for AV events localization and classification. A particular strength of our system is its capability to deal with asynchronous AV events for which typical visual and audio cues appear at different time instants. The proposed experiments have demonstrated the merits of our approach compared to several benchmark methods but have also shown that a more accurate audio temporal modeling would be needed to better cope with situations where the visual modality is unavailable.

Elaborating on temporal modeling more generally, we note that in the discussed instantiation of our framework, each proposal (in both the modalities) is considered independent of the others. Modeling temporal dependency could help identify repeating structures in specific cases such as warning sounds. Incorporating multi-scale analysis could provide better representation for short and long duration audio cues. Similarly, for the image sequence, analyzing temporal structure could reveal important discriminatory motion patterns.

A convenient attribute of our framework is that it can be used in a plug-and-play fashion *i.e.* the component design can be flexibly altered. Of particular interest to us is the proposal design that can be made task-specific. For example, to perform

[8] Localization examples with audio can be found at https://youtu.be/C-jrZ9SDMDY

(a) Train

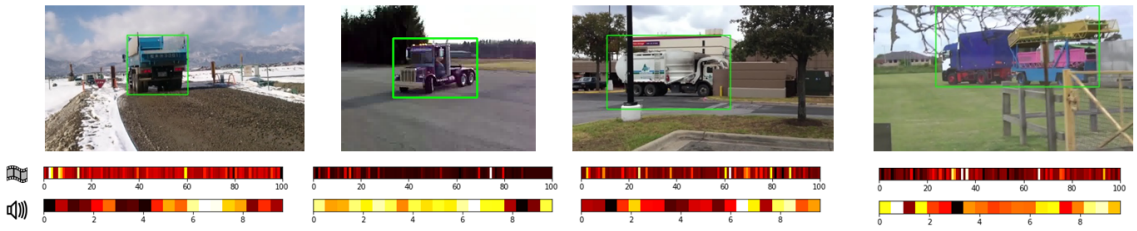


(b) Bicycle

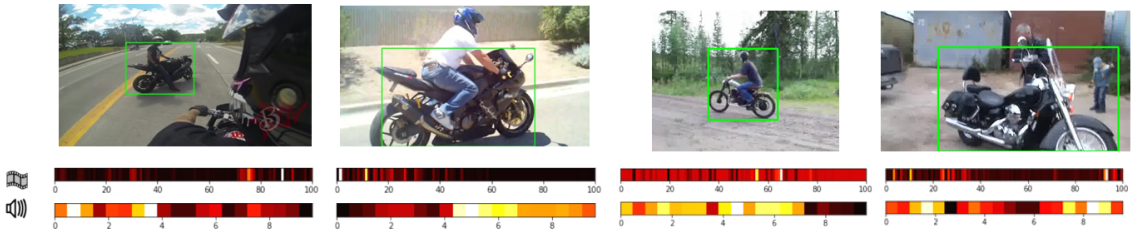
(c) Car

Figure 5.5. Examples of localization on video frames for a few categories from the test data. The localization results are shown in green. Below each image we display the scaled region proposal (top) and audio segment scores for labels referred to in the caption. The visual heatmap is a concatenation of proposals from all the sub-sampled frames, arranged in temporal order.

video object segmentation and tracking, one can incorporate spatio-temporal tube proposals. To this end, it will be interesting to explore the use of recently introduced tube proposal networks (Vu et al., 2018) - a fully convolutional network (Long et al., 2015) that takes as input a set of consecutive frames and outputs class-agnostic tube proposals. However, the method requires object presence or absence supervision for



(a) Truck



(b) Motorcycle

(c) Failure cases

Figure 5.6. Examples of visual localization (continued). The last row shows typical failure cases (in red) which include multiple instance grouping and focus on discriminative regions

training and due to their large number proposals it only considers those corresponding to uniform linear motion. These considerations will need to be appropriately revised for incorporation within our framework.

In the next chapter, we explore one such possibility with the audio modality where we modify the proposals for robust classification and incorporating source separation in noisy scenarios.

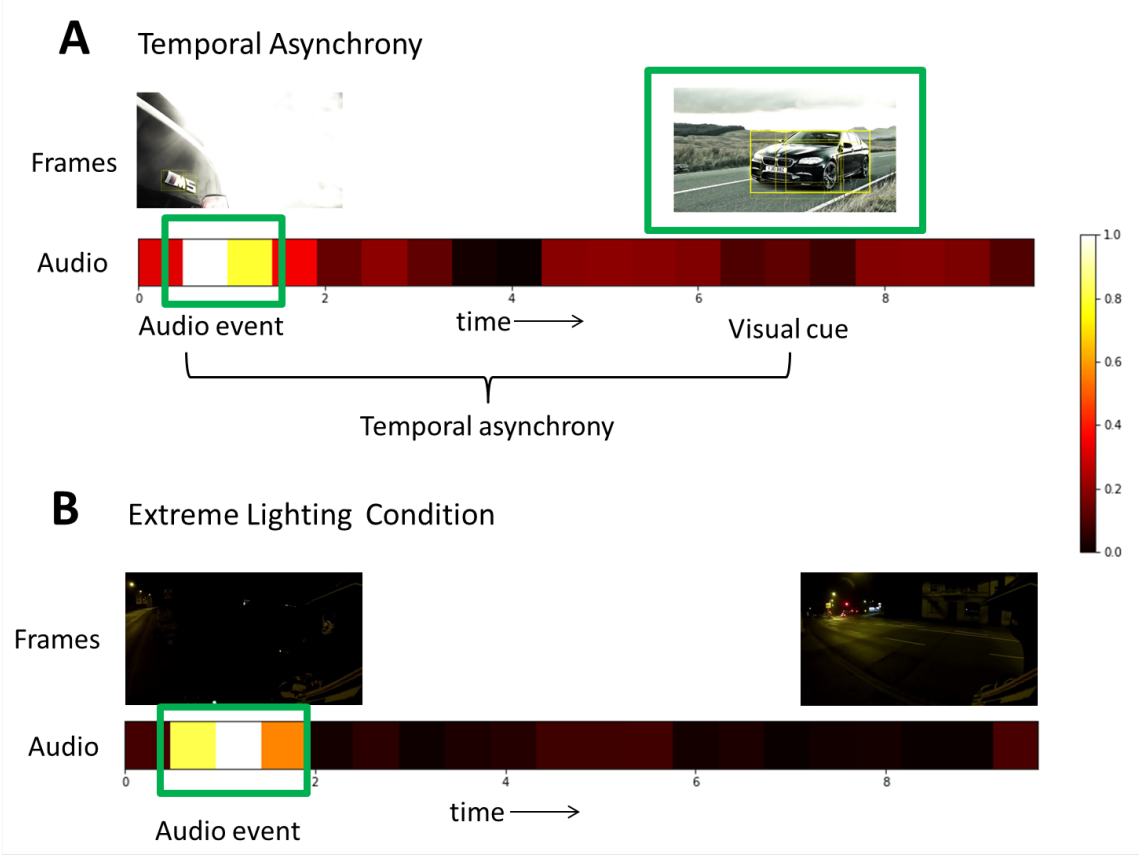


Figure 5.7. Qualitative results for unsynchronized AV events. For both the cases A and B, the heatmap at the bottom denotes audio localization over segments for the class under consideration. For heatmap display, the audio localization vector has been scaled to lie between [0,1]. The top row depicts video frames roughly aligned to the audio temporal axis. (A) Top: Here we show a video where the visual object of interest appears after the audio event. This is a 'car' video from the validation split. The video frames show bounding boxes where edge opacity is controlled by the box's detection score. In other words, higher score implies better visibility (B) Bottom: This is a case from the evaluation data where due to lighting conditions, the visual object is not visible. However the system correctly localizes in audio and predicts the 'motorcycle' class.

6

Robust AV event classification and audio source separation using weak supervision

Synopsis

In this chapter, we build upon our AV 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.

6.1 Introduction

In our efforts to build a unified framework to deal with the challenging problem of automatic AV scene analysis, we presented a first system tackling event identification and AV object localization in the previous chapter. Continuing to build upon that study, in this chapter we focus on making event/object classification robust to noisy/multisource acoustic environments and incorporating the ability to enhance or

separate the object in the audio modality. Specifically, while our earlier approach yields promising results, its audio proposal design has two shortcomings 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 limitations, we propose to generate audio proposals using NMF (Lee and Seung, 2001).

As already discussed, NMF is a popular unsupervised audio decomposition method that has been successfully utilized in various source separation systems (Virtanen, 2007; Ozerov and Févotte, 2010; Durrieu et al., 2011) and as a front-end for audio event detection systems (Heittola et al., 2011; Bisot et al., 2017). The obtained part-based decomposition is analogous to breaking up an image into constituent object regions (Ren and Malik, 2003). 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 the visual modality to perform audio source separation (Gao et al., 2018). There are three key differences with our 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. Another related audio-only approach uses global average pooling and its variants, prevalent in early studies for visual object localization using CNNs (Zhou et al., 2016), to perform source separation and classification from weak labels (Kong et al., 2018). Other recent approaches for deep learning based vision-guided audio source separation methods utilize ground-truth source masks for training (Zhao et al., 2018; Ephrat et al., 2018). It is worth noting that our 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.

Novelty. We show how our deep MIL framework can be flexibly used to robustly

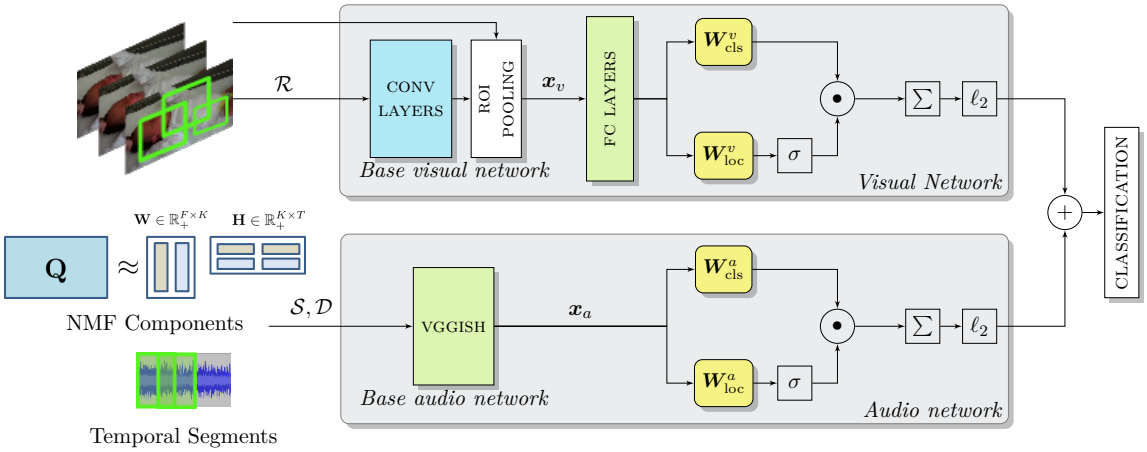


Figure 6.1. Proposed approach: Given a video, we consider the depicted pipeline to go from audio and visual proposals to localization and classification. For the visual modality box proposals are considered, while for audio temporal segments and/or NMF component proposals are utilized. Weights for each module are either trained from scratch (in yellow), fine-tuned (in green) or frozen (in blue) during training.

perform several AV scene understanding tasks using just weak labels. In particular, in addition to temporal audio proposals, 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.

Chapter outline. In Section 6.2 we detail various modules of the proposed approach and source enhancement strategy. This is followed by qualitative and quantitative experimental results on classification, audio source enhancement and visual localization tasks in Section 6.3.

6.2 Proposed approach

The proposed approach is depicted in Fig. 6.1. The problem is formulated within the deep MIL framework discussed in Chapter 5 (Sec. 5.3). We only make changes to the audio branch, keeping all the other components intact. In what follows, we discuss these modifications and for completeness briefly recall the design of other components.

6.2.1 System details

Visual Proposals. We opt for the same visual sub-module as in the previous chapter. Briefly, bounding box proposals are extracted using EdgeBoxes (Zitnick and Dollár, 2014), which are passed through a CNN with RoI pooling layer to obtain fixed length feature vectors, $\boldsymbol{x}_v$.

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, $\mathcal{S} = \{s_1, s_2, \ldots, s_T\}$, as done previously. 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.
2. **NMF Component Proposals (NCP)**: Using NMF (see equation 2.10) we decompose audio magnitude spectrogram $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ consisting of F frequency bins and N short-time Fourier transform (STFT) frames into characteristic audio spectral patterns $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ and their temporal activations $\mathbf{H} \in \mathbb{R}_+^{K \times N}$, respectively. Here K is the total number of spectral patterns.

   We now apply soft mask based filtering as discussed previously in section 2.5, 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 pattern and activation vectors corresponding to the k^{th} component, respectively. The primary reason for performing NMF decomposition is the hope that each of the K tracks would represent a part of just one source.

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 $t \in [1, T]$. 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 `vggish` network (Hershey et al., 2017) for base audio feature extraction. We fine-tune all its layers during training.

Other blocks. The proposal scoring, fusion and classification blocks are left unchanged from our previous study. 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 passing features of each proposal through two-stream scoring architecture (Bilen and Vedaldi, 2016), the ℓ_2 normalized scores are added and used for training with multi-label hinge loss given by Equation 5.2. Both audio and visual sub-networks are trained jointly.

6.2.2 Source enhancement

As noted earlier, a by-product of training the proposed system 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

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

- We apply min-max scaling between [0,1]:

$$\alpha'_k = \frac{\alpha_k - \alpha^l}{\alpha^u - \alpha^l}, \text{ where } \alpha^l = \min_{k'}(\alpha'_k), \ \alpha^u = \max_{k'}(\alpha'_k)$$

- This is followed by soft mask based source and noise spectrogram reconstruction using complex-valued mixture STFT $\mathbf{X}$. 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:

$$\mathbf{S} = \frac{\sum_k \alpha'_k \mathbf{w}_k \mathbf{h}_k}{\mathbf{WH}} \mathbf{X} \tag{6.1}$$

$$\mathbf{N} = \frac{\sum_k (1 - \alpha'_k) \mathbf{w}_k \mathbf{h}_k}{\mathbf{WH}} \mathbf{X} \tag{6.2}$$

  Here $\mathbf{S}$ and $\mathbf{N}$ are the estimates of source of interest and of background noise, respectively. These can be converted back to the time domain using inverse STFT.

6.3 Experiments

6.3.1 Setup

Dataset. We use Kinetics-Instruments (KI), a subset of the Kinetics dataset (Kay et al., 2017) that contains 10-s 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. The details regarding different classes and the number of samples in each is provided in appendix B. 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. 6.1, 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.

While systems using only TSP give state-of-the-art results (Parekh et al., 2018), they 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 (Wang and Plumbley, 2006; Févotte et al., 2018): We use the class labels to train separate dictionaries $\mathbf{W}$ of size $K = 100$ for each music instrument with stochastic mini-batch updates. A supervised NMF variant (as discussed in section 2.5) where the background basis vectors estimated along with activation matrix during test time, while holding the pre–learnt music instrument dictionary (corresponding to the label) fixed.
- NMF Mel-Clustering (Spiertz and Gnann, 2009): 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.[1]

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 $M_{\text{img}} = 10$. 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 (Girshick,

[1] http://www.ient.rwth-aachen.de/cms/dafx09/

2015). 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`.[2] 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 960-ms window with a 480-ms 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 (Lee and Seung, 2001). As stated in Sec. 6.2.1, 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.

Testing Protocol

- *Classification*: Kinetics-Instruments is a multi-class dataset. Hence, we consider $\text{argmax}_c\, s_c$ of the score vector s to be the predicted class and report the overall accuracy computed as the percentage of correct predictions.
- *Source enhancement*: 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 (Stowell et al., 2015).[3] 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.

[2] https://github.com/tensorflow/models/tree/master/research/audioset

[3] For these experiments, we (randomly) chose scene files with suffix 06. The file was appropriately trimmed to match original audio length.

6.3.2 Classification results

In Table 6.1 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 of matrices **W** and **H**. 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.

Table 6.1. Classification results on KI test set. Here, (d) adds the classification scores of systems (a) and (b) at test time [resp. for (h)]

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

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. We also see that using NCP in conjunction with TSP results in a noticable improvement over using just TSP for the audio-only systems. In comparison, this relative difference is negligible for multimodal 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 original audio segments, not split by NMF, could often be easier to classify than individual components.

Table 6.2. Classification accuracy on KI dataset for different levels of noise in the test audio

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

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 6.2. We observe a clear improvement even for the multimodal system when used with NCPs.

6.3.3 Source enhancement results

Following the testing protocol stated in Sec. 6.3.1, we report, in Table 6.3, average Source to Distortion Ratio (SDR) (Vincent et al., 2006) 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 some adaptive thresholding for best results. In other words, as the component scores vary for each example, a single threshold for all cases proves to be sub-optimal. It is worth noting that 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. The training data contains noise in the form of multiple, possibly overlapping instrument and other environmental sounds.

6.3.4 Visual localization examples

We present some promising visual localization examples in Fig. 6.2. The second row of the figure illustrates typical failure cases resulting from multiple object instances, occlusion/object size and intra class variations, respectively. Incorrect localizations due to visual clutter can possibly be remedied at the cost of increasing computational load by considering larger number of proposals or changing their generation method all together (He et al., 2017). Other examples and supplementary material are available on our companion website.[4]

Table 6.3. Average SDR over mixtures created by combining clean instrument examples with environmental scenes.

System	Label Known	Label Unknown
Supervised NMF	2.3	–
NMF Mel-Clustering	–	**4.3**
V + A (NCP), soft	3.3	3.3
V + A (NCP), $\tau = 0.1$	**3.8**	3.9
V + A (NCP), $\tau = 0.2$	3.6	3.6
V + A (NCP, TSP), soft	2.1	2.2

6.4 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

[4] https://perso.telecom-paristech.fr/sparekh/ile2019_supp.html

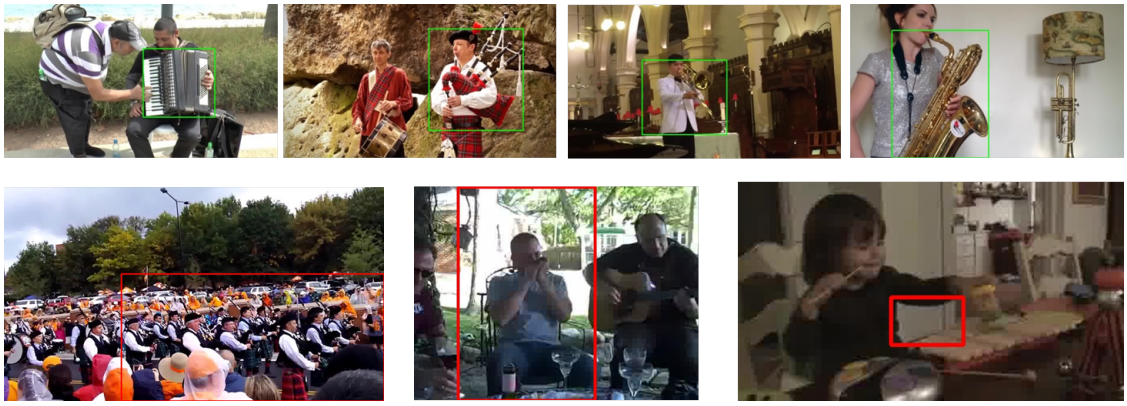


Figure 6.2. Visual localization examples. Top row: correct localization for different instruments (left to right: accordion, bagpipes, trombone and saxophone) from the test set. Max. scoring bounding box shown in green. Bottom row: typical failure cases (in red) such as multiple object instance grouping (bagpipes), small, occluded object (harmonica) and intra-class variation (xylophone) and visual clutter.

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.

While we show encouraging qualitative visual localization results, we do not evaluate for this task quantitatively, in both current and the previous chapter, due to lack of ground-truth in the used datasets. We discuss some strategies in the next chapter to overcome this limitation in the future.

Two research directions worth exploring for further improving source enhancement and the system's performance as a whole are: (i) devising adaptive thresholding strategies for source reconstruction (ii) joint training of NMF or NMF-like deep decompositions with rest of the network such as variational autoencoders, as done in concurrent work (Karamatlı et al., 2018).

7

Conclusion and perspectives

In this chapter we summarize the contributions of this thesis and provide our perspective on future research avenues in light of recent developments in the area of audio-visual analysis.

7.1 Summary of contributions

Taking inspiration from psychology and neuroscience and building upon a rich set of algorithmic investigations, we first identified two broad kinds of joint audio-visual (AV) processing mechanisms for single microphone and camera recordings, namely motion–audio correlation and appearance–audio co-occurrence. The following novel techniques were proposed and explored under each heading, with the goal of performing one or more scene understanding tasks such as event classification, audio source separation, visual object and audio event localization more robustly.

Motion–Audio Correlation

We began by demonstrating the usefulness of jointly processing motion and audio for sound source disambiguation within the popular non–negative matrix factorization (NMF) framework. The novelty lies in our motion-coupled NMF formulation where

the sound producing motion is tied to NMF audio activations through a soft ℓ_1 penalty in the standard Kullback-Leibler divergence cost function. We showed the method's efficacy on string trio recordings using instrument player's motion capture data.

Subsequently, we presented an algorithm to visually assist audio source separation regressing motion from audio. Specifically, we hypothesized that the sound-producing motion's velocity vector can be approximated using a few NMF audio activations – those that correspond to the source. We also suggested an original joint formulation of the problem and derived multiplicative updates. These methods alleviate several shortcomings of the soft–coupling approach referred to earlier, including applicability to generic videos. These constributions are described in chapters 3 and 4.

Appearance–Audio Co-occurrence

Next, we focused on large–scale weakly supervised learning to leverage co-occurrence of audio and appearance characteristics of objects. Our main contributions can be summarized as follows: (1) we presented a new multimodal framework to jointly classify videos and localize both audio and visual cues responsible for this classification; (2) our approach, by construction, allowed dealing with difficult cases when those cues are not synchronized in time; (3) the system's performance was validated, both quantitatively and qualitatively over a weakly-labeled video dataset containing vehicle and warning sound events. State-of-the-art performance was achieved on the task of event classification. We also showed, through a careful analysis of each sub-module, the useful complementary information held in each modality. Qualitative localization results confirmed our technique's ability to identify event-specific AV cues.

Finally, we posited the use of NMF-based audio proposals for allowing our deep multiple instance learning framework to perform audio source separation and robust classification in noisy scenarios. Several experiments were presented on a large dataset of musical instrument videos to substantiate our case.

7.2 Future perspectives

We discuss several future research avenues below.

Linking motion and audio

In chapters 3 and 4 we introduced NMF-based formulations with hand-crafted motion representations. A promising next step would be to build an end-to-end system that automatically extracts and models complex non-linear relationships between motion and audio using techniques such as deep CCA (Andrew et al., 2013). With particular focus on string instruments, a curious follow-up question is: how subtler visual cues such as bow orientation and articulations made using the violin hand (for *e.g.* hand vibrato) can be used for injecting finer details about the produced sound in source separation algorithms?

Another possible extension will be to employ deep generative models for audio decomposition, building upon recent success of variational autoencoders (Karamatlı et al., 2018) and deep unrolling of NMF iterations (Wisdom et al., 2017).

Linking appearance and audio

An interesting future work would be to determine ways of fusing audio and visual modalities at early or intermediate stages within the framework discussed in chapter 5. It will be equally important to extensively explore the use of attention modules (Bahdanau et al., 2014) at different stages – for *e.g.* attending to features of one modality using the other, also called multimodal attention (Lu et al., 2016) or automatically focusing on one of the modalities during classification due to noise in the other. While our preliminary experiments for the latter did not yield better results than the current ones, it is certainly a direction worth delving into deeper.

Combining motion, appearance and audio

A natural next step towards building a more robust, complete system is to put together motion, appearance and audio. A principled way of combining them would

be to learn disentangled audio and visual representations by taking into account motion–audio and appearance–audio relations, as discussed in this thesis.

Another exploratory idea that pertains to the foregoing and current point is the construction of multimodal proposals *i.e.* audio-visual parts that potentially represent different objects in a recording. These could either be fed to a system as input or generated as an intermediate or final output. While their efficacy over current methods is an open question, it would certainly be closer to the human experience of a multimodal object. Bipartite graph clustering for audio-visual signals could be useful in this regard (Ye et al., 2012).

Datasets

Datasets posed a recurring challenge during the course of this thesis. Though available in abundance, the data is largely noisy and unannotated. This is problematic for both training and evaluation. While a long-term effort, there is a need to systematically create a taxonomy, clean and annotate this data, even if only partially, for better performance, benchmarking and progress in this field. AudioSet (Gemmeke et al., 2017) is a commendable start in this direction. Designing efficient audio-visual annotation tools, possibly by extending existing softwares like VATIC (Vondrick et al., 2012), will be of great support in crowd–sourcing the annotation effort. A little explored alternative is the use of synthetic data (Zhang et al., 2017). On the bright side, the described dataset limitation has allowed development of several inspiring non-supervised training procedures, for instance, auxiliary tasks (Arandjelović and Zisserman, 2017a; Owens et al., 2016b).

Audio-visual synthesis

An interesting direction is to use AV relations for synthesis, both as an end goal and a tool for analysis. Here we refer to synthesis of both sounds and visuals. Generating audio for image sequences could happen at different spatial scales - from actions and interactions between two surfaces to generic street or market scenes. Not only is the end result interesting for artists/animators but the process of achieving it requires good understanding of how and which sounds are produced by objects,

materials and actions, learnt as latent representations (Zhang et al., 2017). This is true other way around too *i.e.* generating visuals from audio. Some preliminary work goes in these directions for speaking face (Suwajanakorn et al., 2017) and pose (Shlizerman et al., 2018) and cross-modal synthesis (Chen et al., 2017).

Appendix A

Derivation of Joint NMF-Sparse NNLS algorithm

Derivation of Joint NMF-Sparse NNLS algorithm introduced in chapter 4 for cross-modal regression.

Notations

- $\mathbf{V} \approx \mathbf{WH}$, where $\mathbf{W} = (w_{fk})_{f,k} \in \mathbb{R}_+^{F \times K}$ and $\mathbf{H} = (h_{kn})_{k,n} \in \mathbb{R}_+^{K \times N}$ are interpreted as the nonnegative audio spectral patterns and their activation matrices respectively.
- $\mathbf{M} \in \mathbb{R}_+^{N \times C}$: Velocity matrix with each velocity vector arranged into columns as $[m_1\ m_2 \dots m_C]$
- $\mathbf{A} \in \mathbb{R}_+^{K \times C}$: Nonnegative weight vector for taking linear combinations of $\mathbf{H}$, with each column denoted by α_c where $c \in \{1, C\}$

Details

We formulate the following cost function as:

$$C(\mathbf{W},\mathbf{H},\mathbf{A}) \quad = D_{KL}(\mathbf{V}|\mathbf{WH}) + \lambda\|\mathbf{M}-\mathbf{H}^\intercal\mathbf{A}\|_2^2 + \mu\|\mathbf{A}\|_1$$

Since it is trivial to minimize the cost function we constrain the columns of $\mathbf{W}$ to have unit norm i.e. we construct $\widetilde{\mathbf{W}} = \left[\frac{w_1}{\|w_1\|} \ \frac{w_2}{\|w_2\|} \dots \frac{w_K}{\|w_K\|}\right]$ and incorporate this into the cost function as:

$$\underset{\mathbf{W},\mathbf{H},\mathbf{A}}{\text{minimize}} \quad \underbrace{D_{KL}(\mathbf{V}|\widetilde{\mathbf{W}}\mathbf{H})}_{\text{Audio Factorization}} + \underbrace{\frac{\lambda}{2}\|\mathbf{M}-\mathbf{H}^\intercal\mathbf{A}\|_2^2 + \mu\|\mathbf{A}\|_1}_{\text{Sparse NNLS}} \tag{A.1}$$

For convenience we will use $\mathbf{\Lambda} = \widetilde{\mathbf{W}}\mathbf{H}$.

Update for H. Referring to the factorization part of the cost function in equation A.1 as C_{NMF} and the sparse regression part as C_{SLS}, the derivative with respect to $\mathbf{H}$ gives us the following negative $[.]_-$ and positive $[.]_+$ components:

$$\begin{aligned}
[\nabla_{\mathbf{H}} C_{\text{NMF}}]_+ &= \widetilde{\mathbf{W}}^\intercal\mathbf{1} \\
[\nabla_{\mathbf{H}} C_{\text{NMF}}]_- &= \widetilde{\mathbf{W}}^\intercal\left(\mathbf{V}\odot\mathbf{\Lambda}^{-1}\right) \\
[\nabla_{\mathbf{H}} C_{\text{SLS}}]_+ &= \mathbf{A}\mathbf{A}^\intercal\mathbf{H} \\
[\nabla_{\mathbf{H}} C_{\text{SLS}}]_- &= \mathbf{A}\mathbf{M}^\intercal
\end{aligned} \tag{A.2}$$

Using the heuristic rule explained in section 2.5 (equation 2.15), we can write the update as:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\widetilde{\mathbf{W}}^\intercal\left(\mathbf{V}\odot\mathbf{\Lambda}^{-1}\right) + \lambda\mathbf{A}\mathbf{M}^\intercal}{\widetilde{\mathbf{W}}^\intercal\mathbf{1} + \lambda\mathbf{A}\mathbf{A}^\intercal\mathbf{H}} \tag{A.3}$$

Update for W. For this we follow (Le Roux et al., 2015). The update can be written as:

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{(\mathbf{\Lambda}^{-1}\odot\mathbf{V})\mathbf{H}^\intercal + \widetilde{\mathbf{W}} \odot \left(\mathbf{1}(\widetilde{\mathbf{W}}\odot(\mathbf{1}\mathbf{H}^\intercal))\right)}{\mathbf{1}\mathbf{H}^\intercal + \widetilde{\mathbf{W}} \odot \left(\mathbf{1}(\widetilde{\mathbf{W}}\odot((\mathbf{\Lambda}^{-1}\odot\mathbf{V})\mathbf{H}^\intercal))\right)} \tag{A.4}$$

Update for A. Similar to updates for $\mathbf{H}$, for $\mathbf{A}$ they are easily derived by using C_{SLS} and applying equation 2.15:

$$\mathbf{A} \leftarrow \mathbf{A} \odot \frac{\lambda \mathbf{H}\mathbf{M}}{\lambda \mathbf{H}\mathbf{H}^{\intercal}\mathbf{A} + \mu} \tag{A.5}$$

Appendix B

Dataset details

I. DCASE Smart Cars Challenge Data (Mesaros et al., 2017) - Chapter 5

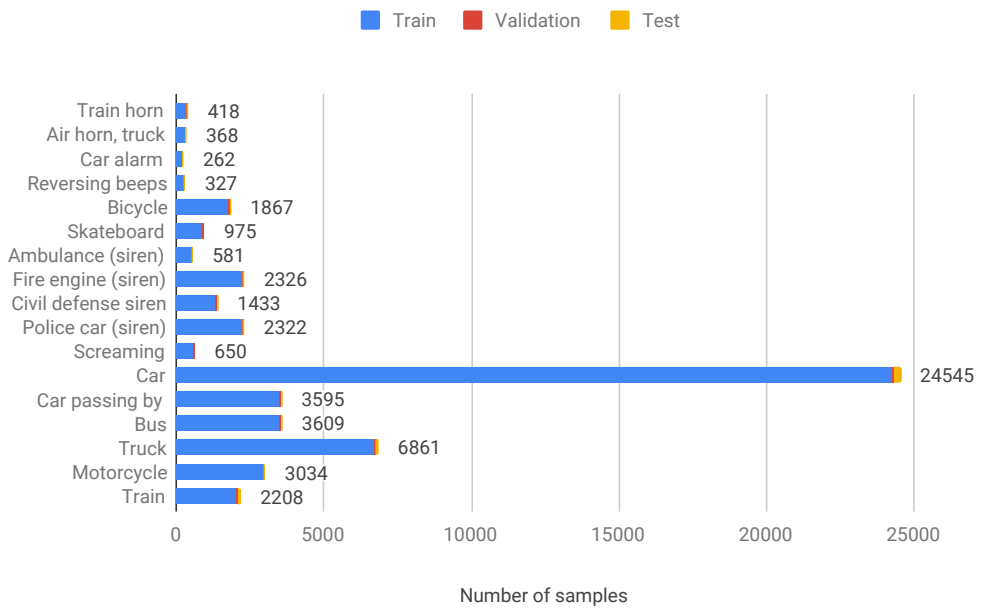


Figure B.1. This highly imbalanced dataset contains 17 different classes shown on the y–axis with the total number of samples indicated by the number at the end of each bar.

This dataset was first introduced for the DCASE smart cars challenge in 2017 (Mesaros et al., 2017).[1] It is a subset of AudioSet (Gemmeke et al., 2017) that consists of weakly labeled YouTube videos (*i.e.* with global video-level event labels) of vehicle and warning sounds. For the purpose of this challenge a few samples were manually annotated with temporal audio event labels. (see Fig. B.1 for more details)

II. Kinetics–Instrument Data (Kay et al., 2017) - Chapter 6

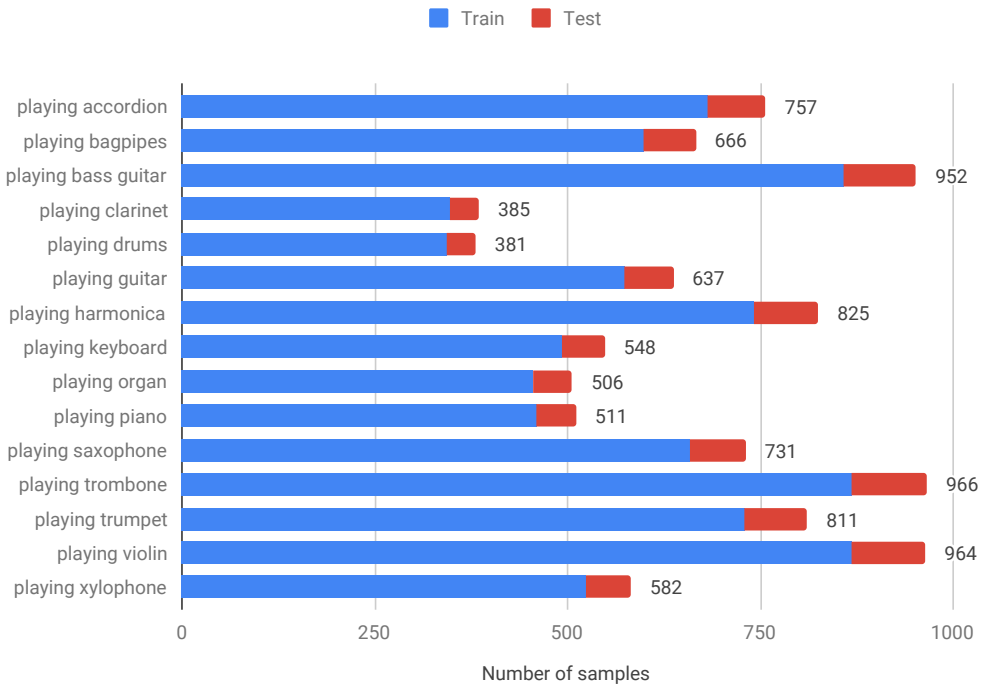


Figure B.2. This dataset contains 15 different instrument classes shown on the y–axis with the total number of samples indicated by the number at the end of each bar.

This subset, taken from the Kinetics action data (Kay et al., 2017) comprises of 15 instrument classes as seen in Fig. B.2. The parent dataset [2] was originally compiled

[1] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-large-scale-sound-event-detection
[2] https://deepmind.com/research/open-source/open-source-datasets/kinetics/

with the aim of human action classification and encompasses human–object/human–human interactions. It comes with video–level action category labels.

Acronyms

A

AV audio-visual

C

CCA canonical correlation analysis

CNN convolutional neural network

D

DCASE Detection and Classification of Acoustic Scenes and Events

K

KL Kullback-Leibler

M

MIL multiple instance learning

N

NMF non-negative matrix factorization

NNLS non-negative least squares

R

RNN recurrent neural network

S

STFT short–time Fourier transform

SVM support vector machine

Bibliography

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, A. P., Toderici, G., Varadarajan, B., and Vijayanarasimhan, S. (2016). Youtube-8M: A large-scale video classification benchmark. In *arXiv:1609.08675*.

Adavanne, S., Pertilä, P., and Virtanen, T. (2017). Sound event detection using spatial features and convolutional recurrent neural network. In *ICASSP*, pages 771–775. IEEE.

Alayrac, J.-B., Bojanowski, P., Agrawal, N., Sivic, J., Laptev, I., and Lacoste-Julien, S. (2016). Unsupervised learning from narrated instruction videos. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (CVPR) (CVPR)*, pages 4575–4583.

Amores, J. (2013). Multiple instance classification: Review, taxonomy and comparative study. *Artificial intelligence*, 201:81–105.

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). Deep canonical correlation analysis. In *Proc. of International Conference on Machine Learning*, pages 1247–1255.

Arandjelović, R. and Zisserman, A. (2017a). Look, listen and learn. In *IEEE International Conference on Computer Vision*.

Arandjelović, R. and Zisserman, A. (2017b). Objects that sound. *CoRR*, abs/1712.06651.

Atilgan, H., Town, S. M., Wood, K. C., Jones, G. P., Maddox, R. K., Lee, A. K., and Bizley, J. K. (2018). Integration of visual information in auditory cortex promotes auditory scene analysis through multisensory binding. *Neuron*, 97(3):640–655.

Aytar, Y., Vondrick, C., and Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900.

Aytar, Y., Vondrick, C., and Torralba, A. (2017). See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Barzelay, Z. and Schechner, Y. Y. (2007). Harmony in motion. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Becker, S. and Hinton, G. E. (1992). Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161.

Bießmann, F., Meinecke, F. C., Gretton, A., Rauch, A., Rainer, G., Logothetis, N. K., and Müller, K.-R. (2010). Temporal kernel cca and its application in multimodal neuronal data analysis. *Mach Learn*, 79(1-2):5–27.

Bilen, H., Namboodiri, V. P., and Van Gool, L. J. (2014a). Object and action classification with latent window parameters. *International Journal of Computer Vision*, 106(3):237–251.

Bilen, H., Pedersoli, M., and Tuytelaars, T. (2014b). Weakly supervised object detection with posterior regularization. In *Proceedings BMVC 2014*, pages 1–12.

Bilen, H. and Vedaldi, A. (2016). Weakly supervised deep detection networks. In *CVPR*, pages 2846–2854.

Bisot, V., Essid, S., and Richard, G. (2017). Overlapping sound event detection with supervised nonnegative matrix factorization. In *ICASSP*, pages 31–35. IEEE.

Bizley, J. K., Jones, G. P., and Town, S. M. (2016). Where are multisensory signals combined for perceptual decision-making? *Current opinion in neurobiology*, 40:31–37.

Boutsidis, C. and Gallopoulos, E. (2008). Svd based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362.

Bredin, H. and Chollet, G. (2006). Measuring audio and visual speech synchrony: methods and applications. *IET International Conference on Visual Information Engineering (VIE 2006)*, pages 255–260.

Bredin, H. and Chollet, G. (2007). Audiovisual speech synchrony measure: application to biometrics. *EURASIP Journal on Applied Signal Processing*, 2007(1):179–179.

Casanovas, A., Monaci, G., Vandergheynst, P., and Gribonval, R. (2010). Blind audiovisual source separation based on sparse redundant representations. *IEEE Transactions on Multimedia*, 12(5):358–371.

Chang, S.-F., Ellis, D., Jiang, W., Lee, K., Yanagawa, A., Loui, A. C., and Luo, J. (2007). Large-scale multimodal semantic concept detection for consumer video. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 255–264. ACM.

Chen, J., Mukai, T., Takeuchi, Y., Matsumoto, T., Kudo, H., Yamamura, T., and Ohnishi, N. (2002). Relating audio-visual events caused by multiple movements: in the case of entire object movement. In *Information Fusion, 2002. Proceedings of the Fifth International Conference on*, volume 1, pages 213–219. IEEE.

Chen, L., Srivastava, S., Duan, Z., and Xu, C. (2017). Deep cross-modal audio-visual generation. In *Proc. of Thematic Workshops of ACM Multimedia*, pages 349–357. ACM.

Cinbis, R. G., Verbeek, J., and Schmid, C. (2017). Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):189–203.

Comon, P. and Jutten, C. (2010). *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press.

Davis, S. B. and Mermelstein, P. (1990). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *Readings in speech recognition*, pages 65–74. Elsevier.

de Sa, V. R. (1994). Learning classification with unlabeled data. In *Advances in neural information processing systems*, pages 112–119.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

Deselaers, T., Alexe, B., and Ferrari, V. (2010). Localizing objects while learning their appearance. In *European conference on computer vision*, pages 452–466. Springer.

Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997a). Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71.

Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997b). Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71.

Dong, L. (2006). *A comparison of multi-instance learning algorithms*. PhD thesis, The University of Waikato.

Duong, N. Q. K., Ozerov, A., Chevallier, L., and Sirot, J. (2014). An interactive audio source separation framework based on non-negative matrix factorization. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1567–1571. IEEE.

Durrieu, J.-L., David, B., and Richard, G. (2011). A musically motivated mid-level representation for pitch estimation and musical audio source separation. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1180–1191.

Eggert, J. and Korner, E. (2004). Sparse coding and nmf. In *International Joint Conference on Neural Networks*, volume 4, pages 2529–2533. IEEE.

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T., and Rubinstein, M. (2018). Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.*, 37(4):112:1–112:11.

Essid, S. and Févotte, C. (2013). Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring. *IEEE Transactions on Multimedia*, 15(2):415–425.

Févotte, C. and Idier, J. (2011). Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456.

Févotte, C., Vincent, E., and Ozerov, A. (2018). Single-channel audio source separation with NMF: divergences, constraints and algorithms. In *Audio Source Separation*, pages 1–24. Springer.

Fisher, J., Darrell, T., Freeman, W. T., Viola, P., and Fisher III, J. W. (2001). Learning Joint Statistical Models for Audio-Visual Fusion and Segregation. In *Advances in Neural Information Processing Systems*, number Ml, pages 772–778. Citeseer.

Fitzgerald, D., Cranitch, M., and Coyle, E. (2009). Using tensor factorisation models to separate drums from polyphonic music. In *Proc Int Conf Digit Audio Eff*.

Fritsch, J. and Plumbley, M. D. (2013). Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 888–891.

Gao, R., Feris, R., and Grauman, K. (2018). Learning to separate object sounds by watching unlabeled video. In *ECCV*.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 776–780. IEEE.

Girshick, R. (2015). Fast R-CNN. In *ICCV*, pages 1440–1448. IEEE.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

Gkioxari, G., Girshick, R., and Malik, J. (2015). Contextual action recognition with r* cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1080–1088.

Goecke, R. and Millar, J. B. (2003). Statistical Analysis of the Relationship between Audio and Video Speech Parameters for Australian English. In *Proc ISCA Tutor Res Workshop Audit-Vis Speech Process*, pages 133–138.

Guo, X., Uhlich, S., and Mitsufuji, Y. (2015). NMF-based blind source separation using a linear predictive coding error clustering criterion. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 261–265.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical correlation analysis: an overview with application to learning methods. *Neural Comput*, 16(12):2639–2664.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916.

Heittola, T., Mesaros, A., Virtanen, T., and Eronen, A. (2011). Sound event detection in multisource environments using source separation. In *Machine Listening in Multisource Environments*.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., et al. (2017). CNN architectures for large-scale audio classification. In *ICASSP*, pages 131–135. IEEE.

Hosang, J., Benenson, R., and Schiele, B. (2014). How good are detection proposals, really? In *25th British Machine Vision Conference*, pages 1–12. BMVA Press.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3 - 4):321–377.

Huang, P.-S., Kim, M., Hasegawa-Johnson, M., and Smaragdis, P. (2014). Deep learning for monaural speech separation. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1562–1566.

Hunter, D. R. and Lange, K. (2004). A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37.

Ilse, M., Tomczak, J. M., and Welling, M. (2018). Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*.

Izadinia, H., Saleemi, I., and Shah, M. (2013). Multimodal analysis for identification and segmentation of moving-sounding objects. *IEEE Transactions on Multimedia*, 15(2):378–390.

Jaiswal, R., FitzGerald, D., Barry, D., Coyle, E., and Rickard, S. (2011). Clustering NMF basis functions using shifted NMF for monaural sound source separation. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 245–248.

Jhuo, I.-H., Ye, G., Gao, S., Liu, D., Jiang, Y.-G., Lee, D., and Chang, S.-F. (2014). Discovering joint audio–visual codewords for video event detection. *Mach Vision Appl*, 25(1):33–47.

Jiang, W., Cotton, C., Chang, S. F., Ellis, D., and Loui, A. (2009a). Short-term audiovisual atoms for generic video concept classification. In *Proc ACM Int Conf Multimed*, pages 5–14. ACM.

Jiang, W., Cotton, C., Chang, S. F., Ellis, D., and Loui, A. (2009b). Short-term audiovisual atoms for generic video concept classification. In *Proceedings of the 17th ACM International Conference on Multimedia*, pages 5–14. ACM.

Jiang, W. and Loui, A. C. (2011). Audio-visual grouplet: temporal audio-visual interactions for general video concept classification. In *Proc ACM Int Conf Multimed*, pages 123–132, Scottsdale, USA.

Jiang, Y.-G., Bhattacharya, S., Chang, S.-F., and Shah, M. (2013). High-level event recognition in unconstrained videos. *International journal of multimedia information retrieval*, 2(2):73–101.

Jiang, Y.-G., Wu, Z., Wang, J., Xue, X., and Chang, S.-F. (2018). Exploiting feature and class relationships in video categorization with regularized deep neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(2):352–364.

Joder, C., Essid, S., and Richard, G. (2008). Temporal Integration for Audio Classification with Application to Musical Instrument Classification. *IEEE Trans Audio Speech Lang Process*.

Kantorov, V., Oquab, M., Cho, M., and Laptev, I. (2016). Contextlocnet: Context-aware deep network models for weakly supervised localization. In *European Conference on Computer Vision*, pages 350–365. Springer.

Karamatlı, E., Cemgil, A. T., and Kırbız, S. (2018). Weak label supervision for monaural source separation using non-negative denoising variational autoencoders. *arXiv preprint arXiv:1810.13104*.

Karras, T., Aila, T., Laine, S., Herva, A., and Lehtinen, J. (2017). Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):94.

Kay, J. (1992). Feature discovery under contextual supervision using mutual information. In *Proc Int Jt Conf Neural Netw*, volume 4, pages 79–84 vol.4.

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

Keuper, M., Andres, B., and Brox, T. (2015). Motion trajectory segmentation via minimum cost multicuts. In *Proc. of IEEE International Conference on Computer Vision (CVPR)*, pages 3271–3279.

Kidron, E., Schechner, Y., and Elad, M. (2005). Pixels that sound. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 88–95 vol. 1.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kolesnikov, A. and Lampert, C. H. (2016). Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision*, pages 695–711. Springer.

Kong, Q., Xu, Y., Wang, W., and Plumbley, M. D. (2018). A joint separation-classification model for sound event detection of weakly labelled data. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 321–325. IEEE.

Kraus, O. Z., Ba, J. L., and Frey, B. J. (2016). Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59.

Krishnan, L., Elhilali, M., and Shamma, S. (2014). Segregating complex sound sources through temporal coherence. *PLoS computational biology*, 10(12):e1003985.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kubovy, M. and Schutz, M. (2010). Audio-visual objects. *Review of Philosophy and Psychology*, 1(1):41–61.

Kumar, A., Khadkevich, M., and Fugen, C. (2017). Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes. *arXiv preprint arXiv:1711.01369*.

Kumar, A. and Raj, B. (2016). Audio event detection using weakly labeled data. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 1038–1047. ACM.

Kumar, M. P., Packer, B., and Koller, D. (2010). Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, pages 1189–1197.

Le Magoarou, L., Ozerov, A., and Duong, N. Q. K. (2015). Text-informed audio source separation. example-based approach using non-negative matrix partial co-factorization. *Journal of Signal Processing Systems*, 79(2):117–131.

Le Roux, J., Weninger, F., and Hershey, J. R. (2015). Sparse NMF–half-baked or well done? *Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023*.

Lee, D., Lee, S., Han, Y., and Lee, K. (2017). Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input. Technical report, DCASE2017 Challenge.

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562.

Li, B., Duan, Z., and Sharma, G. (2016a). Associating players to sound sources in musical performance videos. *Late Breaking Demo, Intl. Soc. for Music Info. Retrieval (ISMIR)*.

Li, B., Liu, X., Dinesh, K., Duan, Z., and Sharma, G. (2016b). Creating a musical performance dataset for multimodal music analysis: Challenges, insights, and applications. *arXiv preprint arXiv:1612.08727*.

Li, D., Dimitrova, N., Li, M., and Sethi, I. (2003). Multimedia content processing through cross-modal association. In *Proc ACM Int Conf Multimed*, Berkeley, CA, USA.

Liutkus, A. and Badeau, R. (2015). Generalized wiener filtering with fractional power spectrograms. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 266–270. IEEE.

Liutkus, A., Durrieu, J.-L., Daudet, L., and Richard, G. (2013). An overview of informed audio source separation. In *14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (CVPR) (CVPR)*, pages 3431–3440.

Lu, J., Yang, J., Batra, D., and Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297.

Maestre, E. (2009). *Modeling instrumental gestures: an analysis/synthesis framework for violin bowing*. Phd thesis, Universitat Pompeu Fabra.

Maragos, P., Gros, P., Katsamanis, A., and Papandreou, G. (2008). Cross-modal integration for performance improving in multimedia: a review. In *Multimodal processing and interaction*, pages 1–46. Springer.

Marchini, M. (2014). *Analysis of Ensemble Expressive Performance in String Quartets: a Statistical and Machine Learning Approach*. PhD thesis, Univesitat Pompeu Fabra.

Marchini, M., Ramirez, R., Papiotis, P., and Maestre, E. (2014). The sense of ensemble: a machine learning approach to expressive performance modelling in string quartets. *Journal of New Music Research*, 43(3):303–317.

Mesaros, A., Heittola, T., Dikmen, O., and Virtanen, T. (2015). Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations. In *ICASSP*, pages 151–155. IEEE.

Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vincent, E., Raj, B., and Virtanen, T. (2017). DCASE2017 challenge setup: Tasks, datasets and baseline system. In *Proc. of Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, pages 85–92.

Mesaros, A., Heittola, T., and Virtanen, T. (2016). Metrics for polyphonic sound event detection. *Applied Sciences*, 6(6):162.

Mysore, G. J. and Smaragdis, P. (2011). A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17–20. IEEE.

Nakadai, K., Hidai, K.-i., Okuno, H. G., and Kitano, H. (2002). Real-time speaker localization and speech separation by audio-visual integration. In *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*, volume 1, pages 1043–1049. IEEE.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *Proc. of International Conference on Machine Learning*, pages 689–696.

Oneata, D., Revaud, J., Verbeek, J., and Schmid, C. (2014). Spatio-temporal object detection proposals. In *European conference on computer vision*, pages 737–752. Springer.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2015). Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694.

Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E. H., and Freeman, W. T. (2016a). Visually indicated sounds. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2405–2413.

Owens, A., Wu, J., McDermott, J. H., Freeman, W. T., and Torralba, A. (2016b). Ambient sound provides supervision for visual learning. In *Proc. of European Conference on Computer Vision*, pages 801–816. Springer.

Ozerov, A. and Févotte, C. (2010). Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):550–563.

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

Ramakrishnan, S. (2018). *Cryptographic and Information Security Approaches for Images and Videos*. CRC Press.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Ren, X. and Malik, J. (2003). Learning a classification model for segmentation. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*.

Rivet, B., Girin, L., and Jutten, C. (2007). Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):96–108.

Salamon, J., McFee, B., and Li, P. (2017). DCASE 2017 submission: Multiple instance learning for sound event detection. Technical report, DCASE2017 Challenge.

Sedighin, F., Babaie-Zadeh, M., Rivet, B., and Jutten, C. (2016). Two multimodal approaches for single microphone source separation. In *EUSIPCO*.

Seichepine, N., Essid, S., Fevotte, C., and Cappe, O. (2014a). Soft nonnegative matrix co-factorization. *IEEE Trans Signal Process*, PP(99).

Seichepine, N., Essid, S., Fevotte, C., and Cappe, O. (2014b). Soft nonnegative matrix co-factorization. *IEEE Transactions on Signal Processing*, PP(99).

Serizel, R., Bisot, V., Essid, S., and Richard, G. (2018). Acoustic features for environmental sound analysis. In *Computational Analysis of Sound Scenes and Events*, pages 71–101. Springer.

Shlizerman, E., Dery, L., Schoen, H., and Kemelmacher-Shlizerman, I. (2018). Audio to body dynamics. In *Proc. CVPR*.

Sigg, C., Fischer, B., Ommer, B., Roth, V., and Buhmann, J. (2007). Nonnegative CCA for Audiovisual Source Separation. In *IEEE Workshop on Machine Learning for Signal Processing*, pages 253–258. IEEE.

Smaragdis, P. and Casey, M. (2003). Audio/visual independent components. In *Proc. of ICA*, pages 709–714. Citeseer.

Smaragdis, P. and Mysore, G. J. (2009). Separation by "humming": user-guided sound extraction from monophonic mixtures. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 69–72.

Smith, L. and Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29.

Song, H. O., Lee, Y. J., Jegelka, S., and Darrell, T. (2014). Weakly-supervised discovery of visual pattern configurations. In *Advances in Neural Information Processing Systems*, pages 1637–1645.

Spiertz, M. and Gnann, V. (2009). Source-filter based clustering for monaural blind source separation. In *in Proceedings of International Conference on Digital Audio Effects DAFx'09*.

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., and Plumbley, M. D. (2015). Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746.

Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):95.

Uijlings, J. R., Van De Sande, K. E., Gevers, T., and Smeulders, A. W. (2013). Selective search for object recognition. *International journal of computer vision*, 104(2):154–171.

Van Gemert, J. C., Jain, M., Gati, E., Snoek, C. G., et al. (2015). Apt: Action localization proposals from dense trajectories. In *Proc. of BMVC*, volume 2, page 4.

Vincent, E., Gribonval, R., and Févotte, C. (2006). Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469.

Vincent, E., Virtanen, T., and Gannot, S. (2018). *Audio source separation and speech enhancement*. John Wiley & Sons.

Virtanen, T. (2007). Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 15(3):1066–1074.

Vondrick, C., Patterson, D., and Ramanan, D. (2012). Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, pages 1–21.

Vu, T.-H., Osokin, A., and Laptev, I. (2018). Tube-cnn: Modeling temporal evolution of appearance for object detection in video. *arXiv preprint arXiv:1812.02619*.

Wang, B. and Plumbley, M. D. (2006). Investigating single-channel audio source separation methods based on non-negative matrix factorization. In *Proc. ICA Research Network International Workshop*, pages 17–20.

Wang, J. and Zucker, J.-D. (2000). Solving multiple-instance problem: a lazy learning approach. In *Proc. of International Conference on Machine Learning*, pages 1119–1126. Morgan Kaufmann Publishers.

Wang, X., Yang, M., Zhu, S., and Lin, Y. (2013). Regionlets for generic object detection. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 17–24. IEEE.

Wisdom, S., Powers, T., Pitton, J., and Atlas, L. (2017). Deep recurrent nmf for speech separation by unfolding iterative thresholding. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2017 IEEE Workshop on*, pages 254–258. IEEE.

Woodhall, W. (2011). *Audio Production and Post-production*. Jones & Bartlett Learning.

Xu, Y., Kong, Q., Wang, W., and Plumbley, M. D. (2017). Surrey-CVSSP system for DCASE2017 challenge task4. Technical report, DCASE2017 Challenge.

Ye, G., Jhuo, I.-H., Liu, D., Jiang, Y.-G., Lee, D., Chang, S.-F., et al. (2012). Joint audio-visual bi-modal codewords for video event detection. In *Proc. of 2nd ACM International Conference on Multimedia Retrieval*, page 39. ACM.

Yokoya, N., Yairi, T., and Iwasaki, A. (2012). Coupled Nonnegative Matrix Factorization Unmixing for Hyperspectral and Multispectral Data Fusion. *IEEE Trans Geosci Remote Sens*, 50(2):528–537.

Yoo, J. and Choi, S. (2011). Matrix co-factorization on compressed sensing. In *Proc Int Joint Conf Artif Intell*.

Yuhas, B. P., Goldstein, M. H., and Sejnowski, T. J. (1989). Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27(11):65–71.

Zhang, C., Platt, J. C., and Viola, P. A. (2006). Multiple instance boosting for object detection. In *Advances in neural information processing systems*, pages 1417–1424.

Zhang, Z., Wu, J., Li, Q., Huang, Z., Traer, J., McDermott, J. H., Tenenbaum, J. B., and Freeman, W. T. (2017). Generative modeling of audible shapes for object perception. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*.

Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., and Torralba, A. (2018). The sound of pixels. In *ECCV*.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2921–2929. IEEE.

Zhou, Y., Xu, Z., Landreth, C., Kalogerakis, E., Maji, S., and Singh, K. (2018). Visemenet: Audio-driven animator-centric speech animation. *ACM Trans. Graph.*, 37(4):161:1–161:10.

Zhuang, X., Zhou, X., Hasegawa-Johnson, M. A., and Huang, T. S. (2010). Real-world acoustic event detection. *Pattern Recognition Letters*, 31(12):1543–1551.

Zitnick, C. L. and Dollár, P. (2014). Edge boxes: Locating object proposals from edges. In *ECCV*, pages 391–405. Springer.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Apprentissage de représentations pour l'analyse robuste de scènes audiovisuelles

Mots clés : Apprentissage automatique, apprentissage de répresentation, traitement du signal audio, vision par ordinateur, analyse de scènes audiovisuelles, séparation de sources

Résumé : L'objectif de cette thèse est de concevoir des algorithmes qui permettent la détection robuste d'objets et d'événements dans des vidéos en s'appuyant sur une analyse conjointe de données audio et visuelle. Ceci est inspiré par la capacité remarquable des humains à intégrer les caractéristiques auditives et visuelles pour améliorer leur compréhension de scénarios bruités. À cette fin, nous nous appuyons sur deux types d'associations naturelles entre les modalités d'enregistrements audiovisuels (réalisés à l'aide d'un seul microphone et d'une seule caméra), à savoir la corrélation mouvement/audio et la co-occurrence apparence/audio.
Dans le premier cas, nous utilisons la séparation de sources audio comme application principale et proposons deux nouvelles méthodes dans le cadre classique de la factorisation par matrices non négatives (NMF). L'idée centrale est d'utiliser la corrélation temporelle entre l'audio et le mouvement pour les objets / actions où le mouvement produisant le son est visible. La première méthode proposée met l'accent sur le couplage flexible entre les représentations audio et de mouvement capturant les variations temporelles, tandis que la seconde repose sur la régression intermodale. Nous avons séparé plusieurs mélanges complexes d'instruments à cordes en leurs sources constituantes en utilisant ces approches.
Pour identifier et extraire de nombreux objets couramment rencontrés, nous exploitons la co-occurrence apparence/audio dans de grands ensembles de données. Ce mécanisme d'association complémentaire est particulièrement utile pour les objets où les corrélations basées sur le mouvement ne sont ni visibles ni disponibles. Le problème est traité dans un contexte faiblement supervisé dans lequel nous proposons un framework d'apprentissage de représentation pour la classification robuste des événements audiovisuels, la localisation des objets visuels, la détection des événements audio et la séparation de sources.
Nous avons testé de manière approfondie les idées proposées sur des ensembles de données publics. Ces expériences permettent de faire un lien avec des phénomènes intuitifs et multimodaux que les humains utilisent dans leur processus de compréhension de scènes audiovisuelles.

Title : Learning representations for robust audio-visual scene analysis

Keywords : Machine learning, representation learning, audio processing, computer vision, audio-visual scene analysis, source separation

Abstract : The goal of this thesis is to design algorithms that enable robust detection of objects and events in videos through joint audio-visual analysis. This is motivated by humans' remarkable ability to meaningfully integrate auditory and visual characteristics for perception in noisy scenarios. To this end, we identify two kinds of natural associations between the modalities in recordings made using a single microphone and camera, namely motion-audio correlation and appearance-audio co-occurrence.
For the former, we use audio source separation as the primary application and propose two novel methods within the popular non-negative matrix factorization framework. The central idea is to utilize the temporal correlation between audio and motion for objects/actions where the sound-producing motion is visible. The first proposed method focuses on soft coupling between audio and motion representations capturing temporal variations, while the second is based on cross-modal regression. We segregate several challenging audio mixtures of string instruments into their constituent sources using these approaches.
To identify and extract many commonly encountered objects, we leverage appearance–audio co-occurrence in large datasets. This complementary association mechanism is particularly useful for objects where motion-based correlations are not visible or available. The problem is dealt with in a weakly-supervised setting wherein we design a representation learning framework for robust AV event classification, visual object localization, audio event detection and source separation.
We extensively test the proposed ideas on publicly available datasets. The experiments demonstrate several intuitive multimodal phenomena that humans utilize on a regular basis for robust scene understanding.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France