

Chapter 9

Multiview Approaches to Event Detection and Scene Analysis

Slim Essid, Sanjeel Parekh, Ngoc Q.K. Duong, Romain Serizel,
Alexey Ozerov, Fabio Antonacci, and Augusto Sarti

Abstract This chapter addresses sound scene and event classification in *multiview* settings, that is, settings where the observations are obtained from multiple sensors, each sensor contributing a particular *view* of the data (e.g., audio microphones, video cameras, etc.). We briefly introduce some of the techniques that can be exploited to effectively combine the data conveyed by the different views under analysis for a better interpretation. We first provide a high-level presentation of generic methods that are particularly relevant in the context of multiview and multimodal sound scene analysis. Then, we more specifically present a selection of techniques used for audiovisual event detection and microphone array-based scene analysis.

Keywords Multimodal scene analysis • Multiview scene analysis • Multichannel audio • Joint audiovisual scene analysis • Representation learning • Data fusion • Matrix factorization • Tensor factorization • Audio source localization and tracking • Audio source separation • Beamforming • Multichannel Wiener filtering

S. Essid (✉)

LTCI, Télécom ParisTech, Université Paris-Saclay, Paris, France
e-mail: slim.essid@telecom-paristech.fr

S. Parekh • Ngoc Q.K. Duong • A. Ozerov
Technicolor, Rennes, France

e-mail: Sanjeel.Parekh@technicolor.com; Quang-Khanh-Ngoc.Duong@technicolor.com;
Alexey.Ozerov@technicolor.com

R. Serizel

Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, France
e-mail: romain.serizel@loria.fr

F. Antonacci • A. Sarti

Politecnico di Milano, Milano, Italy

e-mail: fabio.antonacci@polimi.it; augusto.sarti@polimi.it

9.1 Introduction

We now extend the study of sound scene and event classification to settings where the observations are obtained from multiple sensors, which we refer to as *multiview data*, each sensor contributing a particular *view* of the data. Instances of this include both *multichannel* audio data, as acquired by microphone arrays, or more generally, *multimodal* data, i.e., heterogeneous data that involves two or more *modalities* such as the *audio* or *visual* modalities in video recordings.

Be it for applications in machine perception—at the heart of robots’ and virtual agents’ intelligence systems—or video description—as part of video surveillance or multimedia indexing systems—multiview approaches can lead to a significant boost in performance in challenging real-world situations. Indeed, multiplying the sources of information, through different views, should result in a more robust overall “picture” of the scene being analyzed, where sensors, and consequently views, which are not reliable, e.g., noisy, at a particular time instant, are hopefully backed-up by others. This is, for instance, the case in video recordings where sound-emitting target events are not visible onscreen because of poor lighting conditions or occlusions.

Such an endeavor is actually as promising as challenging, primarily because of the significant increase in the volume of the data to be analyzed, but also owing to the potential heterogeneity of the different streams of information (e.g., audio and visual streams), which additionally may not be perfectly synchronized. Another difficulty is that it is usually not possible to determine which streams are not reliable at every time instant. To see this, consider the scenario of scene analysis using a robot’s sensors. The data views available may then be composed of the multiple audio streams acquired by the robot’s microphone array, as well as RGB and depth-image streams captured by its cameras, possibly along with other signals recorded by inertial measurement units. As the cameras are pointed at an interactant, events of interest may appear only partially in their field of view, and be present in the audio recording only at a very low signal-to-noise ratio. This may be due to background noise (including the robot’s internal noise, the so-called *ego-noise*, typically produced by its cooling fans or its actuators) and the voice of the interactant, or the robot itself, in the foreground.

In this chapter, we briefly introduce some of the techniques that can be exploited to effectively combine the data conveyed by the different views under analysis for a better interpretation. Numerous good surveys have been written on the general topic of multimodal data fusion, notably the paper by Atrey et al. [10] which is quite comprehensive. Therefore, we first provide a high-level presentation of generic methods that are particularly relevant in the context of multiview and multimodal sound scene analysis (Sect. 9.3). It is worth noting that some of the techniques presented have not necessarily yet been considered in the context of scene and event recognition as envisaged in this book. We still briefly cover them in this chapter as it is believed they hold a good potential for such applications. We then more specifically present a selection of techniques used for audiovisual event detection and microphone array-based scene analysis (in Sects. 9.4 and 9.5, respectively).

9.2 Background and Overview

9.2.1 Multiview Architectures

Figure 9.1 depicts an overview of the main fusion strategies that are exploited when analyzing multiview data, namely (1) fusion at the representation or feature level (upper row of the figure), and (2) fusion at the decision-level, usually implying integration of partial classifier-outputs. Each of these methods will be discussed further in Sect. 9.3. In particular, we will focus on a special case of representation/feature-level fusion that is here referred to as *joint subspace learning* where the aim is to learn or non-trivially transform the representations based on inter-relationships across the views.

As previously mentioned, views can be either of the same nature, in which case they are referred to as channels (typically audio channels) each corresponding to a particular microphone, or of different nature as in multimodal scenarios where, for example, some of the views could correspond to different audio channels while others to video images recorded by different cameras.

9.2.2 Visual Features

Since videos are central to the content of this chapter, a short note on commonly employed visual features is in order. Features extracted from visual streams can

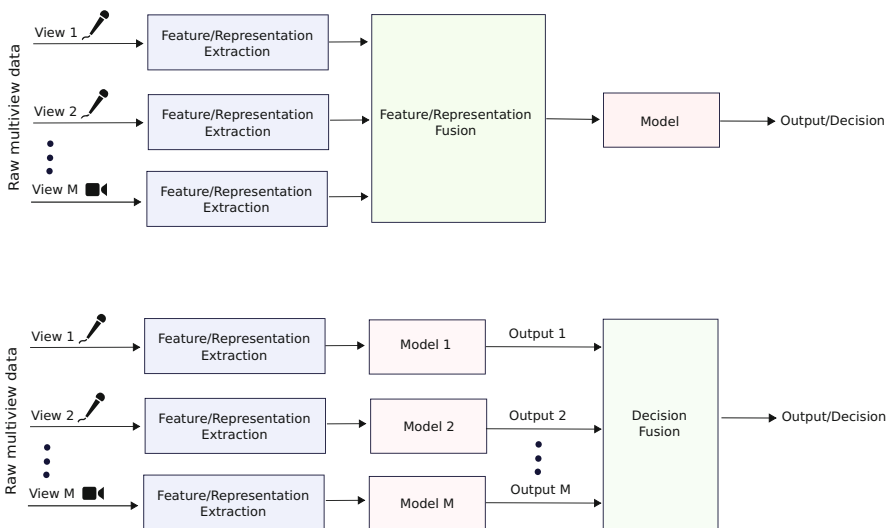


Fig. 9.1 Overview of multiview data analysis approaches

be broadly classified into two categories: appearance-based and motion-based features. For the former, several local and global descriptors representing appearance attributes namely, color, texture, and shape are extracted. While some works utilize the raw pixel data or color histograms, others rely on standard features such as scale-invariant feature transform (SIFT) [96] and histograms of oriented gradients (HOG) [40]. Lately, features extracted from convolutional neural networks have dominated [88].

Motion-based features are typically computed using optical flow or tracking data. It is possible to represent temporal changes of segmented regions, objects, and shapes by calculating velocity and acceleration, i.e., optical flow and its derivative. Other popular features include histograms of optical flow (HOF) [154] and motion boundary histograms (MBH) [154]. As MBH is computed from optical flow derivatives, it is not affected by constant motion. This makes it robust to camera motion. The reader is referred to [79, 101] for an extensive review of visual representations used for multimodal analysis.

In multiview settings temporal synchronization across views is quite challenging. Notably, in the audiovisual case, since the video frame rate, typically around 25–30 frames per second is significantly different from the audio one, features from both modalities must be appropriately sampled for temporal correspondence. Moreover, the natural asynchrony that exists between the two modalities must also be taken into account. This means that cues for an audiovisual event might not appear simultaneously in both modalities.

9.3 General Techniques for Multiview Data Analysis

Generally, the techniques discussed here (in the following two subsections) operate at either the *representation-level* or the *decision-level* as further described in the next sections.

9.3.1 Representation and Feature Integration/Fusion

Representation or feature integration/fusion is the process of combining different types of features or low-level data representations from different views into a common representation (usually to be exploited by a prediction system).

In practice, this can be primarily achieved by concatenating the feature vectors $\mathbf{o}_{m,t}$, extracted from views m , $1 \leq m \leq M$, at the same time positions t , to build integrated feature vectors $\bar{\mathbf{o}}_t = [\mathbf{o}_{1,t}^\top, \dots, \mathbf{o}_{M,t}^\top]^\top$; provided that the data analysis-rate and cross-view synchronization issues have been previously addressed.

However, the dimensionality of the resulting representation is often too high, which has led researchers to resort to dimensionality reduction methods. A common

approach is then to use *feature transform* techniques, possibly principal component analysis (PCA) [31], independent component analysis (ICA) [139], or linear discriminant analysis (LDA) [31] (see Chap. 4, Sect. 4.5.1). An interesting alternative is *feature selection* (see Chap. 4, Sect. 4.5.2). In fact, when applied to the feature vectors $\bar{\mathbf{o}}_t$, the selection will hopefully retain a subset of the most “relevant” features across the various views (with respect to a selection criterion).

Nevertheless, in multimodal settings, the previous methods often turn out to be limited owing to the different physical nature of the features to be combined. In particular, the features do not necessarily live in the same metric spaces, and are not necessarily extracted from the same temporal segments. Consequently, there has been a number of works attempting to address these limitations.

An interesting approach, within the framework of multiple kernel learning, consists in considering separate kernels for different features, to build optimal convex combinations of these in order to use them for classification, as done, for example, in [30, 157].

Another approach that is worthy of note is the construction of joint multimodal representations, as done in video analysis applications, where various types of audiovisual representations have been envisaged. Examples include the creation of *audiovisual atoms* [78] or *audiovisual grouplets* [76], both exploiting audiovisual correlations. A joint audiovisual representation may in particular be built using one of the *joint subspace learning* methods described in the following.

9.3.1.1 Feature-Space Transformation

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, $\mathbf{o}_{1,t} \in \mathbb{R}^{J_1}$ and $\mathbf{o}_{2,t} \in \mathbb{R}^{J_2}$ ($t = 1, \dots, N$), are assembled column-wise in matrices $\mathbf{O}_1 \in \mathbb{R}^{J_1 \times N}$ and $\mathbf{O}_2 \in \mathbb{R}^{J_2 \times N}$, 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 $S_{12}(f_1(\mathbf{O}_1), f_2(\mathbf{O}_2))$ is maximized. Various approaches can be described using this same formalism. The advantages of doing so are twofold: (1) it appropriately modifies the feature spaces to uncover relationships between views specified by the measure of dependency, and (2) by projecting data into the same space, dimensionality difference between views is eliminated and direct comparison across views is made possible. Fisher et al. [51] choose the mutual information [35] as a dependency measure and seek single-layer perceptrons f_1 and f_2 projecting the audiovisual feature vectors to a 2-dimensional space. Other more popular approaches (for which closed-form solutions can be found) use linear mappings to project the feature streams:

¹The 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 down-sampling the audio features or up-sampling the video features, or by using temporal integration techniques [80].

- Canonical correlation analysis (CCA), first introduced by Hotelling [67], aims at finding pairs of unit-norm vectors \mathbf{t}_1 and \mathbf{t}_2 such that

$$(\mathbf{t}_1, \mathbf{t}_2) = \arg \max_{(\mathbf{t}_1, \mathbf{t}_2) \in \mathbb{R}^{J_1} \times \mathbb{R}^{J_2}} \text{corr}(\mathbf{t}_1^\top \mathbf{O}_1, \mathbf{t}_2^\top \mathbf{O}_2) \quad (9.1)$$

CCA can be considered equivalent to mutual information maximization for the particular case where the underlying distributions are elliptically symmetric [83]. Several variants have been proposed to incorporate sparsity and non-negativity into the optimization problem to resolve issues with interpretability and ill-posedness, respectively [84, 138]. In the context of multimodal neuronal data analysis, temporal kernel CCA [15] 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:

$$(\mathbf{t}_1, \mathbf{t}_2) = \arg \max_{(\mathbf{t}_1, \mathbf{t}_2) \in \mathbb{R}^{J_1} \times \mathbb{R}^{J_2}} \text{cov}(\mathbf{t}_1^\top \mathbf{O}_1, \mathbf{t}_2^\top \mathbf{O}_2) \quad (9.2)$$

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 [21].

- Yet another configuration known as cross-modal factor analysis (CFA), and found to be more robust than CCA in [92], seeks two matrices \mathbf{T}_1 and \mathbf{T}_2 , such that

$$(\mathbf{T}_1, \mathbf{T}_2) = \arg \max_{(\mathbf{T}_1, \mathbf{T}_2)} (1 - \|\mathbf{T}_1 \mathbf{O}_1 - \mathbf{T}_2 \mathbf{O}_2\|_F^2) = \arg \min_{(\mathbf{T}_1, \mathbf{T}_2)} \|\mathbf{T}_1 \mathbf{O}_1 - \mathbf{T}_2 \mathbf{O}_2\|_F^2 \quad (9.3)$$

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

Note that all the previous techniques can be kernelized to study nonlinear coupling between the modalities considered (see, for instance, [64, 90]).

The interested reader is referred to [64, 67, 92] for further details on these techniques, and to [58] for a comparative study.

9.3.1.2 Multimodal Dictionary Learning

While previous approaches relied on modeling the association between the features across modalities, this class of techniques targets the extraction of meaningful multimodal structures to jointly represent all the modalities. This is useful because feature transformation techniques like CCA impose simplifying assumptions such as linearity and are adversely affected by lack of data. To this end, Monaci et al. [106] propose to learn multimodal dictionaries wherein the dictionary elements are learned using an algorithm that enforces synchrony between modalities and decorrelation between the learned dictionary elements. The learned templates can

then be used for performing various tasks. Monaci et al. improve upon this foundational work by proposing a bimodal matching pursuit algorithm which integrates dictionary learning and coding [107]. The sparse shift-invariant generative model used for the audiovisual case can be given by defining multimodal dictionary elements $\{\phi_d\}_{d=1}^D = (\phi_d^a(t), \phi_d^v(x, y, t))$ consisting of audio, ϕ_d^a , and visual, ϕ_d^v , parts and a spatio-temporal shift operator $T_{(pqr)}\phi_d = (\phi_d^a(t-r), \phi_d^v(x-p, y-q, t-r))$ such that the multimodal signal s is approximated by the following equation:

$$s \approx \sum_{d=1}^D \sum_{i=1}^{n_d} c_{d_i} T_{(pqr)d_i} \phi_d \quad (9.4)$$

where n_d is the number of instances of ϕ_d and c_{d_i} specifies the weights for AV components of ϕ_d at the i th instance. Several limitations of this approach have been improved upon by proposing a new objective function and algorithm to balance the two modalities, reduce computational complexity, and improve robustness [94].

9.3.1.3 Co-Factorization Techniques

Matrix factorization techniques can be profitably used to extract meaningful representations for the data being analyzed.

When dealing with *multichannel* data—i.e., with data views of the same nature (e.g., multichannel audio or images)—observations from multiple channels may be profitably assembled in *multi-way arrays*, i.e., *tensors*, before being modeled by tensor factorization methods. As for multichannel audio data, a popular approach consists in collecting the spectrograms of signals from different channels (originating from different microphones) in a 3-way tensor, as illustrated in Fig. 9.2, before processing it with the so-called PARAFAC (PARALLEL FACtor analysis) decomposition method, possibly with non-negativity constraints. This can be interpreted as an attempt to explain audio spectra observations \mathbf{v}_{tm} as being linear combinations of elementary spectra \mathbf{w}_k , temporally weighted by activation coefficients o_{kt} up to spatial modulation coefficients q_{mk} .

Such decompositions were found particularly useful in multichannel audio source separation [52, 118]. For more information about tensor factorization methods, we refer the reader to [33, 87, 159].

In contrast to the previous setting, data from different modalities usually live in feature spaces of completely different topology and dimensionality (think of audio as opposed to video), preventing the possibility of “naturally” representing them by the same tensor. In this case, one may resort to the 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 has usually to be characterized through temporal dependencies between the temporal activations in cross-modal correspondence, and unlikely through dependencies between dictionary elements of different modalities.

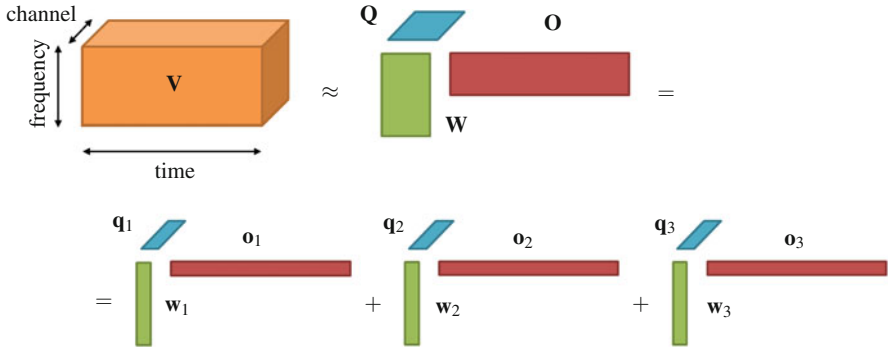


Fig. 9.2 PARAFAC decomposition of multichannel audio spectra

Assuming that appropriate nonnegative features have been extracted at the same rate from the two modalities being analyzed²—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, for the audio and visual data. One may seek a model $(\mathbf{W}_1, \mathbf{W}_2, \mathbf{O})$ such that:

$$\begin{cases} \mathbf{V}_1 \approx \mathbf{W}_1 \mathbf{O} \\ \mathbf{V}_2 \approx \mathbf{W}_2 \mathbf{O} \\ \mathbf{W}_1 \geq 0, \mathbf{W}_2 \geq 0, \mathbf{O} \geq 0; \end{cases} \quad (9.5)$$

in such a way that the temporal activations be the 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., [53, 160, 161]). 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. [134], which merely encourages the temporal activations corresponding to each modality to be close, as opposed to equal, according to:

$$\begin{cases} \mathbf{V}_1 \approx \mathbf{W}_1 \mathbf{O}_1 \\ \mathbf{V}_2 \approx \mathbf{W}_2 \mathbf{O}_2 \\ \mathbf{O}_1 \approx \mathbf{O}_2 \\ \mathbf{W}_1 \geq 0, \mathbf{W}_2 \geq 0, \mathbf{O}_1 \geq 0, \mathbf{O}_2 \geq 0. \end{cases} \quad (9.6)$$

²To 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.

The model (9.6) is estimated by solving the following optimization problem:

$$\begin{cases} \min_{\theta} C_c(\theta) ; \theta \triangleq (\mathbf{W}_1, \mathbf{O}_1, \mathbf{W}_2, \mathbf{O}_2) \\ \mathbf{W}_1 \geq 0, \mathbf{W}_2 \geq 0, \mathbf{O}_1 \geq 0, \mathbf{O}_2 \geq 0 ; \end{cases} \quad (9.7)$$

$$C_c(\theta) \triangleq D_1(\mathbf{V}_1 | \mathbf{W}_1 \mathbf{O}_1) + \gamma D_2(\mathbf{V}_2 | \mathbf{W}_2 \mathbf{O}_2) + \delta P(\mathbf{O}_1, \mathbf{O}_2) ; \quad (9.8)$$

where:

- $D_1(\cdot | \cdot)$ and $D_2(\cdot | \cdot)$ 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;
- $P(\cdot, \cdot)$ is a penalty on the difference between (properly rescaled) activation values occurring at the same instant; they can be either 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 [134] for more details on the algorithms.³ The soft co-factorization scheme has proven effective for multichannel [134] and multimodal audio source separation [120], as well as multimodal speaker diarization [133]. It is believed to be promising for audiovisual event detection tasks.

9.3.1.4 Neural Networks and Deep Learning

Lately, rapid progress in the application of deep learning methods to representation learning has motivated researchers to use them for fusing multiview data [3, 112, 140]. The primary advantage of neural networks is their ability to model very complex nonlinear correlations that exist between multiple views. Early insights into their use for multiview data were provided by Yuhas et al. [163] who trained a network to predict audio using visual input. Subsequently, Cutler et al. [39] proposed to learn audiovisual correlations for the task of speaker detection using a time-delayed neural network (TDNN). Recently, various multimodal autoencoder architectures for learning shared representations have been proposed, even for the case where only a single view is present at training and testing time [112]. Another interesting work extends CCA to learning two deep encodings, one for each view, such that their correlation is maximized [3]. Regularized deep neural networks [158] have also been proposed to construct shared representations taking into account the feature inter-relationships. Each of these methods has been developed independently in different settings. Their application to event analysis and detection still remains to be explored. This is a rapidly growing area of research; we refer the interested reader to [56, 68] for recently proposed multimodal fusion architectures.

³Matlab implementations are available online at <http://plato.telecom-paristech.fr/publi/26108/>.

9.3.2 Decision-Level Integration/Fusion

Decision-level fusion, also known as *late integration* refers to the idea of combining intermediate decisions, i.e., partial classifier-outputs, in order to achieve a more accurate multimodal characterization of a content, an idea which has been explored extensively, under various configurations. This can be seen as a particular case of *ensemble learning* [125] where the base classifiers (to be combined) operate on different views of the data.

Numerous works rely on *majority voting* procedures whereby final global decisions are made based on a weighted sum of individual voters, each typically corresponding to a decision taken on a particular view. The weights are often chosen using either heuristics or trial-and-error procedures (see, for example, [93]). This idea can be better formalized using a Bayesian framework that allows for taking into account the uncertainty about each classifier decisions [71, 103].

9.3.2.1 Probabilistic Combination Rules

When using classifiers providing local probabilistic outputs $p(G_c | \mathbf{o}_{m,t})$ for the t -th observation of the m -th view, $\mathbf{o}_{m,t}$, a simplistic decision strategy assumes feature-vector observations from different views to be independent, and the decision rule consequently takes the form:

$$\hat{G} = \arg \max_c \log[p(G_c | \mathbf{o}_{0,t}, \dots, \mathbf{o}_{M-1,t})] = \arg \max_c \sum_{m=0}^{M-1} \log p(G_c | \mathbf{o}_{m,t}). \quad (9.9)$$

It is worth mentioning that alternative simple combination rules have also been employed that are discussed in Kittler et al. [86].

The previous approach does not allow for incorporating prior knowledge about the dependency structure in the data, in particular the cross-modal and temporal dependencies. To this end, sophisticated dynamic classifiers have been utilized, ranging from variants of (multistream) hidden Markov Models (HMM) [7, 60, 85, 111], through more general dynamic Bayesian networks [32, 59, 109], to even more general graphical models such as conditional random fields (CRF) [20].

9.3.2.2 Neural Networks

Neural networks can also be used for late integration. Some works have utilized them to adaptively learn the weights for fusing multiple classifiers or system outputs [74, 113]. This is typically carried out by training the network to minimize the error between estimated and oracle weights [74]. Besides, in order to take into account the temporal and multiview dependencies, a frequently used strategy is to perform end-to-end training with “fusion” integrated as a layer (usually close to the output

layer) into the architecture [49, 82]. Such methods cannot be termed as late fusion in the conventional sense as feature learning and decision fusion stages are not independent.

9.3.2.3 Other Methods

Another widespread strategy consists in using the monomodal classifier-outputs as features, on the basis of which a new classifier, that is expected to optimally perform the desired multimodal fusion, is learned [13, 156]. Also, solutions to deal with the potential imprecision of some views have been proposed using the *Dempster–Shafer* theory [54].

Finally, it is important to note that the techniques described in this section are not mutually exclusive: in practice one may jointly consider different integration strategies for different features and views (possibly being driven by some expert knowledge), and different analysis time-horizons. This raises the difficult issue of effectively and efficiently exploiting, at the final prediction stage, heterogeneous representations: low-level instantaneous features, possibly over varying time-scales, intermediate prediction results—sometimes seen as outputs of *event* or *concept* detectors—bags-of-words or bags-of-systems extracted over longer texture-windows, *etc.*

9.4 Audiovisual Event Detection

9.4.1 Motivation

The target of audiovisual event detection (AVED) is to detect specific events that occur in an audiovisual recording or real-time stream, and to identify the class of those events. Though the task is more widely addressed through the analysis of the video images, information conveyed by the sound track may become key for a proper detection. Indeed, the visual information may not be sufficient since occlusions may occur and events may be localized in space, hence not visible in the images, given that the camera field of view is necessarily restricted. Also the images may not be usable because of poor lighting conditions, or fast camera motion. AVED then enables a more reliable detection of these events, by combining audio and visual cues.

9.4.1.1 Examples in Video Content Analysis and Indexing

Researchers continue to explore various techniques for improving video content analysis and indexing for better navigation and user experience. In this context,

AV event analysis at various levels of granularity provides useful insights into the composition of such data in terms of objects, activities, and scenes. This not only improves retrieval but also provides a representation closer to our understanding of the physical world. For example, a user could search a database for activity videos such as “dribbling a basketball” or “playing a violin.” Evidently, these are two very distinct tasks where the differences can be readily detected based on auditory and motion information. Moreover, joint analysis could reveal the presence of various objects (e.g., violin, basketball) and also the surroundings (e.g., concert hall, court).

Such an analysis makes object detection and segmentation [72], concept classification [75, 76, 78], scene segmentation and change detection[149], activity analysis, and various other related tasks possible. Several systems submitted to TRECVID⁴ video content analysis tasks of multimedia event detection, story segmentation, and search rely on AV analysis [2, 77, 156].

9.4.1.2 Examples in AV Surveillance and Robot Perception

Video has recently become an increasingly important resource for forensics and surveillance [104, 124]. Video captured by CCTV systems or video recorded from mobile devices (and possibly shared on multimedia platforms) can provide essential clues in solving criminal cases. For example, when considering an investigation about a missing person, video documents can help to localize the missing person or a suspect, providing crucial information about their whereabouts. The analysis of videos linked with a missing person or her/his social network can also help to understand the conditions of the disappearance (was it a kidnapping, a runaway, etc.) and largely influence the investigation.

An investigator looking for a video in a large dataset may want to retrieve information based on the type of scene where the video was recorded or also, at a finer granularity level, based on specific events that occurred during the recording. In addition, the detection of specific events can help to confirm (or deny) the fact that a video was recorded in a particular scene. Some events are indeed representative of particular scenes. For example, train noise in all probability indicates the scene takes place in a train station. Plates and cutlery noises indicate the scene is probably taking place in a restaurant [22]. On the other hand, some events are unlikely to happen in particular scenes. AVED can then help tracking anomalies to detect abnormal events (gunshots, crowd panic, etc.) [97] or to identify a recording scene where information has voluntarily been concealed. This is the case, for example, when a kidnapper sends a ransom video recorded from inside a building but a church bell or a train passing nearby can be heard during the video. This type of information that is not present visually can help to localize the place where the video was recorded [136].

⁴TREC Video Retrieval Evaluation: <http://www-nlpir.nist.gov/projects/trecvid/>.

9.4.2 AV Event Detection Approaches

9.4.2.1 AV Event Detection and Concept Classification

Approaches to AV event detection have been very varied and data dependent. Many works for traditional event detection utilize Markov model variants such as the duration dependent input–output Markov model (DDIOMM)[110], multistream HMM, or coupled HMM [69]. The former uses a decision-level fusion strategy and the latter two do it at an intermediate level. These methods have been shown to perform better than single modality-based approaches with coupled-HMMs being particularly useful for modeling AV asynchrony.

Specifically, with regard to event detection in surveillance videos, Cristiani et al. [37] propose to use the AV concurrence matrix to identify salient events. The idea is to model the audio/video foreground and construct this matrix based on the assumption that simultaneously occurring AV foreground patterns are likely to be correlated. Joint AV analysis has also been employed extensively for sports video analysis and for broadcast analysis in general. In one approach, several feature detectors are built to encode various characteristics of field sports. Their decisions are then combined using a support vector machine (SVM)[127]. Several approaches for structuring TV news videos have also been proposed.

On the other hand, *joint codebook-based approaches* have been quite popular for the task of multimedia concept classification.⁵ In essence, each element of these multimodal codebooks captures some part of a salient AV event. Work on short-term audiovisual atoms (S-AVA) [78] aims to construct a codebook from multimodal atoms which are a concatenation of features extracted from tracked short-term visual-regions and audio. To tackle the problem of video concept classification, this codebook is built through multiple instance learning. Following this work, AV *grouplets* (AVG) [76] 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. As an alternative, Jhuo et al. [75] 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.

⁵Here the term “concept classification” refers to generic categorization in terms of scene, event, object, or location [78].

9.4.2.2 AV Object Localization and Extraction

AV object localization and extraction refers to the problem of identifying sources visually and/or aurally. This section serves to show how objects responsible for audiovisual events can be extracted from either of the modalities through joint analysis. The general approach is to first associate the two modalities using methods discussed in Sect. 9.3. The parameters learned during the former step can then be utilized for object localization and segmentation (visual part), audio source separation (audio counterpart), or unsupervised AV object extraction in both modalities. We now discuss approaches to each of these application scenarios.

Object localization and segmentation has been a popular research problem in the computer vision community. 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. [51] proposed to use joint statistical modeling to perform this task using mutual information. Izadinia et al. [72] 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 [84, 138]. Effectiveness of CCA can be illustrated with a simple example of a video with a person dribbling a basketball [72] (see Fig. 9.3). Simplifying Izadinia et al.'s [72] visual feature extraction methodology, we compute the optical flow and use mean velocity calculated over 40×40 blocks as the visual representation and mel-spectra as the audio representation. The heat map in Fig. 9.3 shows correlation between each image block and audio. Areas with high correlation correspond to regions with motion. If we instead use a soft co-factorization model [134], it is indeed possible to track the image blocks correlated with the audio in each frame.

Another approach worth mentioning is one that uses Gestalt principles for locating sound sources in videos [105]. Inspired by Gestalt principle of temporal



Fig. 9.3 CCA illustration: heat map showing correlation between video image regions and audio. *Black squares indicate highest correlation*

proximity the authors propose to detect synchronous audiovisual events. A particularly different approach was taken by Casanovas et al. [29] who proposed an audiovisual diffusion coefficient to remove information from video image parts which are not correlated with the audio.

Audio source separation is the audio counterpart of the previously discussed problem. The aim is to extract sound produced by each source using video information. As done for videos, mutual information maximization has been used to perform source separation in a user-assisted fashion by identifying the source spatially. Recent methods perform this within the NMF-based source separation framework [120, 132].

Several other approaches deal with both object segmentation and source separation *together* in a completely unsupervised manner. Work by Barzeley et al. [11] considers onset coincidence to identify AV objects and subsequently perform source separation. A particular limitation of this method is the requirement of setting multiple parameters for optimal performance on each example. Blind AV source separation work has also been attempted using nonnegative CCA [138] and sparse representations [28]. Independent component analysis over concatenated features from both modalities also extracts meaningful audiovisual objects [139]. However its application is limited to static scenes. Finally, multimodal dictionary learning has also been utilized in this context [94].

While the methods discussed in this section have been shown to work well in controlled environments, their performance is expected to degrade in dense audiovisual scenarios. Moreover, they make a simplifying assumption that all the objects are seen onscreen. It must be emphasized that most of these techniques can be considered symmetric, in the sense that they can be applied to tasks in either of the modalities with appropriate representations.

9.5 Microphone Array-Based Sound Scene Analysis

In complex sound scenes the sounds coming from different sources can be overlapping in time and frequency. Single channel processing can discriminate sources based on time or frequency as long as they are separated in either time or frequency. Trying to detect or classify sound events that are overlapping both in time and frequency directly from a single channel signal will generally result in a confusion between events. An alternative approach is to attempt to separate individual events prior to detection or classification. However, trying to separate sounds that are overlapping both in time and frequency with single channel techniques is known to be problematic and will inevitably introduce a loss of information resulting in a degradation of the subsequent detection and classification performance. *Microphone arrays* enable the usage of multichannel techniques that exploit not only temporal and spectral diversity between sources but also spatial information about their location.

Historically microphone arrays were composed of a set of microphones placed along a straight line (with constant spacing between microphones for linear arrays or variable spacing for logarithmic arrays). Less constraint arrays have been used for specific purposes such as spherical and circular arrays and more recently arrays without spatial constraints in the case of wireless acoustic sensor networks became a popular research topic. Some approaches and concepts presented here are applicable only to specific array topologies or at least when the topology is known beforehand (see also below). In this chapter we also assume that the signals coming from different microphones are synchronized at the stage of sampling in order to allow for the exploitation of spatial cues. Readers should keep in mind that at the time of writing of this book, dealing with unsynchronized microphone arrays is still an open research problem.

9.5.1 Spatial Cues Modeling

In order to exploit spatial information about the sound sources, audio scene analysis algorithms usually first model the spatial cues and then estimate the corresponding parameters. Both deterministic and probabilistic modeling of such spatial cues have been widely considered in the literature. The former case usually relies on (a) the *point source* assumption, where sound from a source is assumed to come from a single position, and (b) the *narrowband approximation*, where a mixing process from an audio source to the microphone array is characterized by a mixing frequency dependent vector [98]. Probabilistic modeling is usually applied for reverberated or diffuse sources, where sound from a source may come from many directions due to the reverberation, e.g., source localization [18, 63, 116], separation [46, 73, 99], and beamforming systems [17, 48]. This section will discuss some typical spatial cue models, in both a deterministic and a probabilistic sense, for different audio scene analysis applications.

9.5.1.1 Binaural Approach

Humans generally combine cues from several audiovisual streams to localize sound sources spatially. The main cues for localization in the horizontal hemisphere are related to binaural hearing (relying on the difference between the signal reaching the right ear and the signal reaching the left ear). All these cues are encoded in the so-called interaural transfer function (ITF) that includes the following:

- The **interaural time difference (ITD)** is the difference between the time-of-arrival of a signal at the left ear and the right ear. It is useful to localize sounds based on their onset and at low frequency (below 1.5 kHz) [89] (see also Fig. 9.4a).

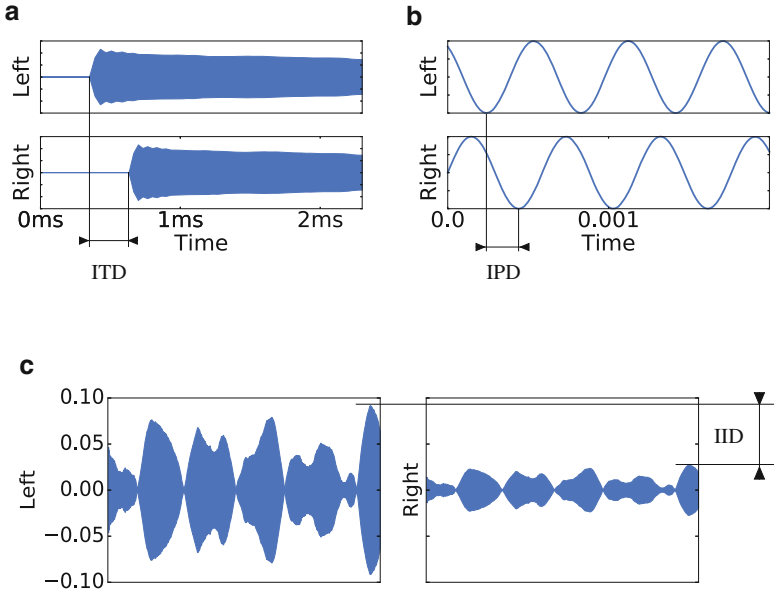


Fig. 9.4 Artificial representation of the binaural cues. (a) ITD, (b) IPD, (c) IID

- The **interaural phase difference (IPD)** is the phase difference between the signal at the left ear and the right ear. It is useful to localize on-going sound as long as the wavelength is larger than the diameter of the head (below 1.5 kHz) [162] (see also Fig. 9.4b);
- The **interaural intensity difference (IID)** is the difference in level between the signal at the left ear and the right ear due to the acoustic shadow produced by the head for sounds above 3 kHz (below the so-called head shadow effect is not present) [108] (see also Fig. 9.4c).

All the concepts mentioned above can be extended to general microphone array setups. The ITD and IPD concepts directly generalize to linear microphone arrays where they relate rather straightforwardly to time difference of arrival (TDOA) and direction of arrival (DOA). In this case, however, the arrays have to be designed carefully to prevent spatial aliasing. The IID concept is less applicable to small linear arrays as it relies on the head shadow effect. Indeed, in small arrays the level difference between the signal impinging two consecutive microphones might not be significant. However, in ad-hoc arrays where the topology is unconstrained, the microphones can be quite far apart and IID can become insightful as well, granted that the microphone positions are known beforehand. These spatial cues are extensively exploited to extract a signal of interest from the mixture using beamforming approaches described in Sect. 9.5.1.2 (for example, the delay-and-sum beamformer directly relies on ITD). Spatial cues can also be used directly for sound source localization (see also Sect. 9.5.2.3) and, by proxy, for source separation (see also Sect. 9.5.2.1) and sound event detection (see also Sect. 9.5.2.2).

9.5.1.2 Beamforming Methods

Fixed beamformers compose a first simple class of multichannel algorithms which can separate signals coming from different directions. A fixed beamformer tries to steer toward the direction from where the desired sound signal comes and to reject signals coming from other directions. The main categories of fixed beamformers include delay-and-sum beamformers, filter-and-sum beamformers [66], superdirective microphone arrays [36], or the original formulation of the minimum variance distortionless beamformer (MVDR) [26].

Adaptive beamformers try to steer toward the direction of the desired sound signal and to adaptively minimize the contributions from the undesired sources coming from other directions. This typically yields a constrained optimization problem. Frost introduced the linearly constrained minimum variance beamformer (LCMV) as an adaptive framework for MVDR [55].

The generalized side lobe canceler (GSC), also known as the Griffiths-Jim beamformer, is an alternative approach to the LCMV where the optimization problem is reformulated as an unconstrained problem [62]. The GSC can be decomposed as a fixed beamformer steering toward the desired source, a blocking matrix, and a multichannel adaptive filter [65].

The multichannel Wiener filters (MWF) represent another class of multichannel signal extraction algorithms which are defined by an unconstrained optimization problem [45]. MWF-based algorithms can be implicitly decomposed into a spatial filter and a spectral filter, and can indeed be considered as beamformers [135]. Besides, a reformulation of MWF allows for explicitly controlling the spectral distortion introduced [45, 135].

9.5.1.3 Nonstationary Gaussian Model

The nonstationary Gaussian framework has emerged in audio source separation [46, 50, 114, 119] as a probabilistic modeling of the reverberated sources. It was then also applied in, e.g., multichannel acoustic echo cancellation [144] and multichannel speech enhancement [145]. In this paradigm, the short-time Fourier transform (STFT) coefficients of the *source images* $\mathbf{c}_j(t, f)$, i.e., the contribution of the j -th source ($1 \leq j \leq J$) at the microphone array, are modeled as a zero-mean Gaussian random vector whose covariance matrix $\widehat{\mathbf{R}}_j(t, f) = \mathbb{E}(\mathbf{c}_j(t, f)\mathbf{c}_j^H(t, f))$ can be factorized as

$$\widehat{\mathbf{R}}_j(t, f) = v_j(t, f)\mathbf{R}_j(t, f), \quad (9.10)$$

where $v_j(t, f)$ are scalar time-varying *variances* encoding the spectro-temporal power of the sources and $\mathbf{R}_j(t, f)$ are $I \times I$ *spatial covariance matrices* encoding their spatial position and spatial width. This model does not rely on the point source assumption nor on the narrowband assumption, hence it appears applicable to

reverberated or diffuse sources. In the general situation where the sound source can be moving, the spatial cues encoded by $\mathbf{R}_j(t, f)$ are time-varying. However, in most cases where the source position is fixed and the reverberation is moderate, the spatial covariance matrices are time-invariant: $\mathbf{R}_j(t, f) = \mathbf{R}_j(f)$. Different possibilities of parameterizing $\mathbf{R}_j(f)$ have been considered in the literature resulting in either the rank-1 or the full-rank matrices, where the later case was shown to be more appropriate for modeling the reverberated and diffuse sources as it accounts directly for the interchannel correlation in the off-diagonal entries of $\mathbf{R}_j(f)$ [46].

9.5.2 Spatial Cues-Based Sound Scene Analysis

This section will discuss the use of spatial cue models presented in the previous section in some specific applications, namely sound source separation, acoustic event detection, and moving sound source localization and tracking.

9.5.2.1 Sound Source Separation

In daily life, recorded sound scenes often result from the superposition of multiple sound sources which prevent both human and machines from well localizing and perceiving the target sound sources. Thus, source separation plays a key role in sound scene analysis, and its goal is to extract the signals of individual sound sources from an observed mixture [98]. It offers many practical applications in, e.g., communication, hearing aids, robotics, and music information retrieval [6, 14, 100, 152].

Most source separation algorithms operate in the time-frequency (T-F) domain with the mixing process formulated as

$$\mathbf{x}(t, f) = \sum_{j=1}^J \mathbf{c}_j(t, f) \quad (9.11)$$

where $\mathbf{x}(t, f) \in \mathbb{C}^{I \times 1}$ denotes the STFT coefficients of the I -channel mixture at T-F point (t, f) , and $\mathbf{c}_j(t, f) \in \mathbb{C}^{I \times 1}$ is the j -th source image. As $\mathbf{c}_j(t, f)$ encodes both spectral information about the sound source itself and the spatial information about the source position, a range of spectral and spatial models has been considered in the literature resulting in various source separation approaches. In the determined case where $I \geq J$, non-Gaussian modeling such as frequency-domain independent component analysis (FDICA) has been well-studied [122, 128]. In the under-determined situation where $I < J$, sparse component analysis (SCA) has been largely investigated [19, 61, 81]. As a specific example of the nonstationary Gaussian modeling presented in Sect. 9.5.1.3, the parameters are usually estimated by the expectation maximization (EM) algorithm derived in either the maximum

likelihood (ML) sense [46] or the maximum a posteriori (MAP) sense [47, 117, 119]. Then source separation is achieved by the multichannel Wiener filtering. Readers are referred to, e.g., [95, 150] for the survey of recent advances on both *blind* scenarios and *informed* scenarios which exploit some prior knowledge about the sources themselves [119] or the mixing process [47] to better guide the source separation.

9.5.2.2 Sound Event Detection

As different sound events usually occur at different spatial locations in the sound scene, spatial cues obtained from microphone array processing intrinsically offer important information for SED. As an example, information about the source directions inferred from the interchannel time differences of arrival (TDOA) was used to help partitioning home environments into several areas containing different types of sound events in [151]. The combination of these spatial features with the classic MFCC was reported to improve the event classification in the experiment. Motivated by binaural processing, in [1] the stereo log-mel-band energy is extracted from stereo recordings to train the neural networks in order to obtain a meaningful cue similarly to the IID.

9.5.2.3 Localization and Tracking of Sound Sources

Sound source localization and tracking are concerned with estimating and following the position of a target source within a sound scene. This active field of research in microphone array processing finds important applications, e.g., in surveillance or video conferencing where the camera should be able to follow the moving speaker, and even can automatically switch the capture to an active sound source in multiple source environments [153]. Spatial cues offered by the multichannel audio capture play a key role in deriving the algorithms.

The problem of acoustic source localization has been a relevant topic in the audio processing literature for the past three decades because of its applicability to a wide range of applications [41, 146]. The most effective solutions rely on the use of spatial distributions of microphones, which sample the sound field at several locations. Spurious events, reverberation, and environmental noise, however, can be a significant cause of localization error. In order to ease the problem, at least for those errors that are contained in a limited number of time frames, source tracking techniques can come in handy, as they are able to perform trajectory regularization, even on the fly. Typical approaches are based on particle [5, 91, 155], Kalman [4, 57], or distributed Kalman filtering [143].

Different methodologies have been developed for the localization of acoustic sources through microphone arrays. Those that gained in popularity are based on measurements of the time delay between array microphones. Working on the time domain is often a suitable choice for wideband signals, and most techniques tend to

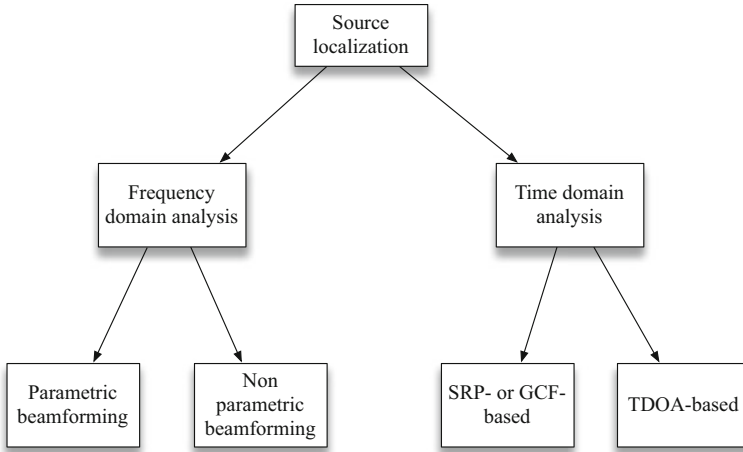


Fig. 9.5 Taxonomy of source localization techniques

rely on the analysis of the generalized cross-correlation (GCC) of the signals [27] and variants thereof. Localization in the frequency domain, however, can be shown to attain good results for narrowband or harmonic sources immersed in a wideband noise and rely on the analysis of the covariance matrix of the array data. A taxonomy of the localization techniques is represented in Fig. 9.5.

Time-Domain Localization

Steered response power (SRP, [42, 43, 102]) and global coherence field (GCF, [115]) proceed through the computation of a coherence function that maps the GCC values at different microphone pairs on the hypothesized source location. A source location estimate is found as the point in space that maximizes the coherence function. In [23] the scenario of multiple sources is accommodated through a two-step procedure that, after localizing the most prominent source, deemphasizes its contribution in the GCC, so that other sources can be localized. These techniques are known for their high level of accuracy, and are suitable for networks of microphone arrays, where synchronization can only be guaranteed between microphones of the same array. One limitation of such solutions is their computational cost, which is proportional to the number of hypothesized source locations. This means that increasing the spatial resolution results in higher computational costs. Some solutions have been proposed in the literature to mitigate this problem. In [165] the authors propose a hierarchical method that begins with a coarser grid, and refines the estimate at different steps by computing the map for finer grids concentrated around the candidate locations estimated at the previous step. In [44] a similar approach is adopted, but a stochastic region contraction strategy is used for going from a coarser to a finer grid. An example of steered response power with stochastic region contraction map is shown in Fig. 9.6.

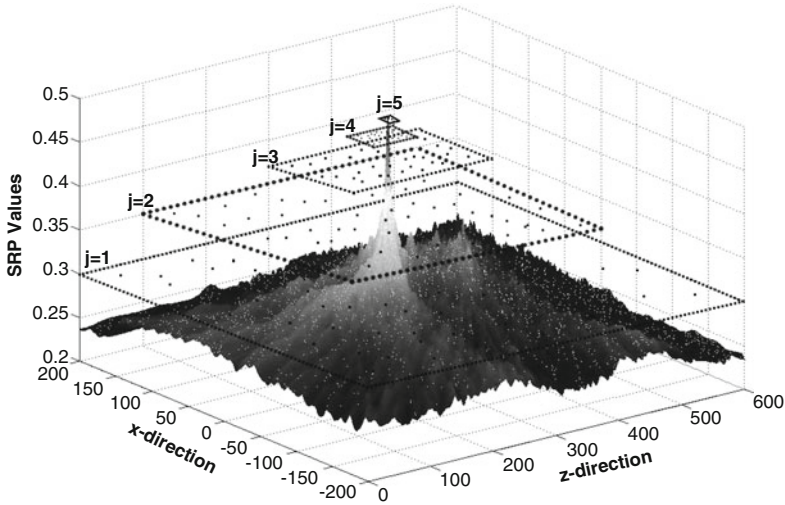


Fig. 9.6 Example of a coherent map using the Steered Response Power with Stochastic Region Contraction technique (SRP-SRC, from [44])

Less cumbersome are the solutions based on the time difference of arrival (TDOA), which is estimated as the time lag of the GCC that exhibits the maximum value. The TDOA is then converted into range difference (RD), which measures the difference of the range between the source and the two microphones in the pair. The locus of candidate source locations corresponding to a given TDOA is a branch of hyperbola whose foci are in the microphones locations, and whose aperture is proportional to the measured TDOA. The most straightforward technique for localization consists in intersecting branches of hyperbolas corresponding to the TDOA measurements coming from different pairs of microphones. The cost function that is based on this procedure is strongly nonlinear, which makes the method sensitive to measurement errors. Least squares cost functions provide a good approximation [12, 34, 70, 129]. The main drawback of TDOA-based localization is its sensitivity to outlier measurements. In [24, 25, 130] techniques for removal of the outliers were presented. In particular, the DATEMM algorithm [130] is based on the observation that TDOAs over a closed loop must sum to zero.

Frequency-Domain Localization

Techniques in the frequency domain are based on the observation that different microphones in the array will receive differently delayed replicas of the source signals. This, in the frequency domain, corresponds to a phase offset. For distant sources the phase offset between adjacent microphones is constant throughout the array. Delay-and-sum beamformers compensate the offsets so that the components

related to a direction will sum up coherently and the others will not. The estimation of the direction of arrival (DOA) of the target source proceeds by searching for the direction that maximizes the output energy of the beamformer over a grid of directions [141, Chapter 6]. The most straightforward nonparametric beamformer is the delay and sum, which is known for its low resolution capabilities, making it difficult to distinguish sources that are seen under close angles from the array viewpoint. The minimum variance distortionless response beamformer (MVDR, [26]) partially improves the resolution capabilities. Parametric techniques, among which it is worth mentioning multiple signal classification (MUSIC, [131]), and estimation of the signal parameters through rotational invariance techniques (ESPRIT, [126]) bring improvements in terms of resolution. However, they are known for their sensitivity to noise and reverberation, which tends to introduce spurious localizations. The superdirective data-independent beamformer [16] was shown to partially mitigate this problem. An interesting solution to the sensitivity to reverberation was proposed in [137] for the detection of gunshots using networks of sensors, each equipped with four or more microphones. For each sensor, both DOA and TDOA are measured. Source location is estimated by intersecting the loci of potential source locations (hyperbolas and direction of arrival) for the two kind of measurements from all the sensors. In reverberant conditions and in the presence of interferers, some TDOAs and some DOAs could be related to spurious paths, thus providing multiple estimates of the gunshot location. The actual gunshot location is found as the one that maximizes the number of consistent TDOAs and DOAs.

It is important to notice that TDOA-based and frequency-domain source localization techniques require the synchronization of the microphones within the array. This, in fact, becomes an issue when multiple independent small arrays are deployed in different locations. In [25] the authors propose a technique for the localization without requiring a preliminary synchronization of the arrays by including the time offsets between the arrays into the unknowns, along with the location of the source. Another important issue is the self-calibration of the array, i.e., the estimation of the mutual relative positions of the microphones [38, 147]. The widespread diffusion of mobile phones and devices equipped with one or more microphones enables the implementation of a wireless acoustic sensor network in seconds, for goals ranging from teleconferencing to security. In this context, however, both calibration and synchronization are needed before normal operation [123].

Acoustic Source Tracking

Independently of the adopted localization method, reverberation and interferers could introduce spurious localizations. The goal of source tracking is to alleviate the influence of outliers. The idea behind tracking is that measurements related to the actual source must follow a dynamical model whereas those related to spurious sources must not [155]. Another goal that can be pursued with tracking systems is that of fusing information coming from both audio and visual localization systems [9, 142]. Several solutions have been presented in the literature. The Kalman filter

[57] is a linear system characterized by two equations. The state equation models the evolution of the state of a system (location and speed of the source) from one time frame to the next one. The observation equation links the state variables with the observable measurements. The goal of the Kalman filter is to estimate the current state from the knowledge of time series of the observations.

Recently, distributed Kalman filters have been used, which enable the tracking of acoustic sources also in the case of distributed array networks [164], without requiring that all nodes communicate the whole state of the system.

Inherent assumptions that lie in the use of the Kalman filter are the linearity and Gaussianity of measurement and state vectors. In order to gain in robustness against the nonlinearity, the use of the extended Kalman filter has been proposed [142], which linearizes the nonlinear system around the working point. In order to gain in robustness against non-Gaussian conditions, however, one has to resort to a different modeling of the source dynamics. In recent years particle filter gained interest in the source localization community due to the fact that it is suitable also to perform tracking in nonlinear non-Gaussian systems and, more in general, for its higher performance [155]. Particle filtering [8] assumes that both state and measurement vectors are known in a probabilistic form. Once a new measurement vector is available, the likelihood function of the current observation from a given state is sampled through particles. Each particle is assigned a weight, which determines its relevance in the likelihood function. Only relevant particles will be propagated to the next step. The source location is determined as the centroid of the set of particles. An example of tracking of one, two, or three acoustic sources on a given trajectory for DOA measurements is shown in Fig. 9.7.

In audio surveillance contexts, it is important to enable localization also when multiple sources are active at any time, with a small convergence time when acoustic sources alternate. This is important, for example, in events that involve multiple acoustic sources (brawls, people yelling, etc.). In recent years, swarm particle filtering has shown to address this scenario particularly well [121]. It is based on the idea that the propagation of each particle to the next step is determined not only by the previous history of the particle itself, but also by the particle that exhibits the best likelihood at the current time instant. Consequently, the overall behavior of the systems resembles that of a bird flock, rapidly moving toward the active source. An example of behavior of swarm particle filtering is shown in Fig. 9.8. Here two sets of particles at four consecutive time frames estimate the location of a source using particle filtering (PF) and swarm particle filtering (Swarm). The two sets are initialized identically. It is possible to notice that after four steps, the swarm particles cluster around the source location, while the PF is still converging.

9.6 Conclusion and Outlook

Multichannel and multimodal data settings represent opportunities to address complex real-world scene and event classification problems in a more effective manner. The availability of concurrent, hence potentially complementary streams of data is amenable to a more robust analysis, by effectively combining them, using

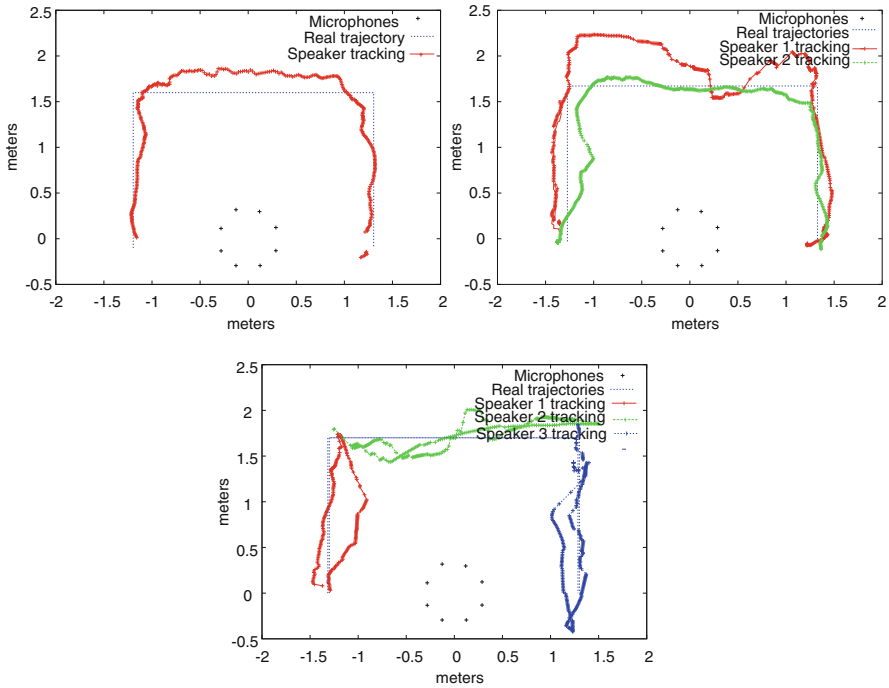
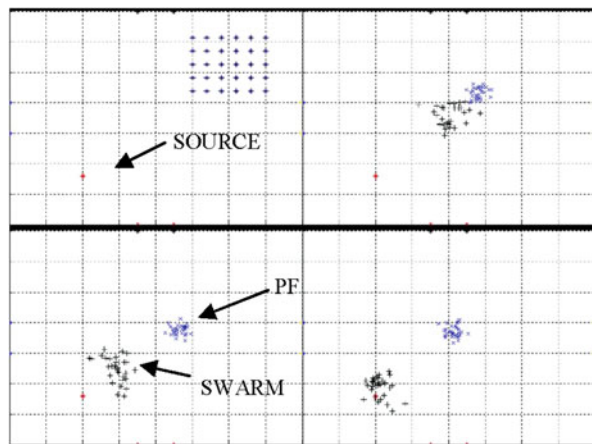


Fig. 9.7 Example of tracking of one, two, or three sources over a prescribed trajectory (from [148])

Fig. 9.8 Example of behavior of two sets of particles propagated using particle filtering (PF) and swarm particle filtering (SWARM). The two sets of particles occupy the same location at the first time frame (from [121])



appropriate techniques, be it at the input representation-level, the feature-level, or the decision-level. Successful applications of such techniques have been realized in various multichannel audio and audiovisual scene analysis tasks.

Yet, a number of research questions remain open in these settings. Notably, it is still not clear how to generically detect when some of the data views are temporarily not reliable (typically noisy or out of focus, with respect to the classes of interest) and which strategies should be developed that can efficiently ignore such views and proceed with the classification (or any other similar data processing) using models which were perhaps trained assuming all views are available.

Also, given the complexity of accurately annotating all data views, especially for instantaneous multi-label event classification tasks, that is when multiple events may occur simultaneously, it is important to consider learning methods that can take advantage of very coarse ground-truth labels, which may have been obtained based on just one of the views, without necessarily being relevant for others. An example of this is the “blind” annotation of the audio track of a video (without considering the images) where sound events may not be visible onscreen at the same time stamps. Multiple instance learning and weakly supervised learning techniques may turn out to be effective learning paradigms to address these difficulties.

References

1. Adavanne, S., Parascandolo, G., Pertila, P., Heittola, T., Virtanen, T.: Sound event detection in multichannel audio using spatial and harmonic features. In: Proceedings of the IEEE AASP Chall Detect Classif Acoust Scenes Events (2016)
2. Amir, A., Berg, M., Chang, S.F., Hsu, W., Iyengar, G., Lin, C.Y., Naphade, M., Natsev, A., Neti, C., Nock, H., et al.: Ibm research trecvid-2003 video retrieval system. In: NIST TRECVID-2003 (2003)
3. Andrew, G., Arora, R., Bilmes, J.A., Livescu, K.: Deep canonical correlation analysis. In: Proceedings of the International Conference on Machine Learning (2013)
4. Antonacci, F., Lonoce, D., Motta, M., Sarti, A., Tubaro, S.: Efficient source localization and tracking in reverberant environments using microphone arrays. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 4, pp. iv–1061. IEEE, New York (2005)
5. Antonacci, F., Matteucci, M., Migliore, D., Riva, D., Sarti, A., Tagliasacchi, M., Tubaro, S.: Tracking multiple acoustic sources in reverberant environments using regularized particle filter. In: Proceedings of the International Conference on Digital Signal Processing, pp. 99–102 (2007)
6. Arai, T., Hodoshima, H., Yasu, K.: Using steady-state suppression to improve speech intelligibility in reverberant environments for elderly listeners. *IEEE Trans. Audio Speech Lang. Process.* **18**(7), 1775–1780 (2010)
7. Argones Rúa, E., Bredin, H.H., García Mateo, C., Chollet, G.G., González Jiménez, D.: Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden Markov models. *Pattern Anal. Appl.* **12**(3), 271–284 (2008)
8. Arulampalam, M., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Process.* **50**(2), 174–188 (2002)
9. Asoh, H., Asano, F., Yoshimura, T., Yamamoto, K., Motomura, Y., Ichimura, N., Hara, I., Ogata, J.: An application of a particle filter to Bayesian multiple sound source tracking with audio and video information fusion. In: Proceedings of the Fusion, pp. 805–812. Citeseer (2004)

10. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. *Multimed Syst* **16**(6), 345–379 (2010)
11. Barzelay, Z., Schechner, Y.Y.: Harmony in motion. In: *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1–8 (2007)
12. Beck, A., Stoica, P., Li, J.: Exact and approximate solutions of source localization problems. *IEEE Trans. Signal Process.* **56**(5), 1770–1778 (2008)
13. Benmokhtar, R., Huet, B.: Neural network combining classifier based on Dempster-Shafer theory for semantic indexing in video content. In: *International MultiMedia Modeling Conference (MMM 2007)*, Singapore, 9–12 January 2007. LNCS, vol. 4352/2006, Part II. <http://www.eurecom.fr/publication/2119>
14. Bertin, N., Badeau, R., Vincent, E.: Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Trans. Audio Speech Lang. Process.* **18**(3), 538–549 (2010)
15. Bießmann, F., Meinecke, F.C., Gretton, A., Rauch, A., Rainer, G., Logothetis, N.K., Müller, K.R.: Temporal kernel cca and its application in multimodal neuronal data analysis. *Mach. Learn.* **79**(1–2), 5–27 (2010)
16. Bitzer, J., Simmer, K.U.: Superdirective microphone arrays. In: *Microphone Arrays*, pp. 19–38. Springer, New York (2001)
17. Bitzer, J., Simmer, K.U., Kammeyer, K.D.: Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2965–2968 (1999)
18. Blandin, C., Ozerov, A., Vincent, E.: Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. *Signal Process.* **92**(8), 1950–1960 (2012)
19. Bofill, P., Zibulevsky, M.: Underdetermined blind source separation using sparse representations. *Signal Process.* **81**(11), 2353–2362 (2001)
20. Bousmalis, K., Morency, L.P.: Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In: *International Conference on Automatic Face & Gesture Recognition*, pp. 746–752 (2011)
21. Bredin, H., Chollet, G.: Measuring audio and visual speech synchrony: methods and applications. *Proceedings of the IET International Conference on Visual Information Engineering*, pp. 255–260 (2006)
22. Bregman, A.S.: *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge (1994)
23. Brutti, A., Omologo, M., Svaizer, P.: Localization of multiple speakers based on a two step acoustic map analysis. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4349–4352 (2008)
24. Canclini, A., Antonacci, F., Sarti, A., Tubaro, S.: Acoustic source localization with distributed asynchronous microphone networks. *IEEE Trans. Audio Speech Lang. Process.* **21**(2), 439–443 (2013)
25. Canclini, A., Bestagini, P., Antonacci, F., Compagnoni, M., Sarti, A., Tubaro, S.: A robust and low-complexity source localization algorithm for asynchronous distributed microphone networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **23**(10), 1563–1575 (2015)
26. Capon, J.: High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE* **57**(8), 1408–1418 (1969)
27. Carter, G.C.: Coherence and time delay estimation. *Proc. IEEE* **75**(2), 236–255 (1987)
28. Casanovas, A., Monaci, G., Vanderghenst, P., Gribonval, R.: Blind audiovisual source separation based on sparse redundant representations. *IEEE Trans. Multimed.* **12**(5), 358–371 (2010)
29. Casanovas, A.L., Vanderghenst, P.: Nonlinear video diffusion based on audio-video synchrony. *IEEE Trans. Multimed.*, 2486–2489 (2010). doi:[10.1109/ICASSP.2010.5494896](https://doi.org/10.1109/ICASSP.2010.5494896)
30. Chang, S.F., Ellis, D., Jiang, W., Lee, K., Yanagawa, A., Loui, A.C., Luo, J.: Large-scale multimodal semantic concept detection for consumer video. In: *Proceedings of the International Workshop on Multimedia Information Retrieval, MIR '07*, pp. 255–264. ACM, New York, NY (2007)

31. Chibelushi, C.C., Mason, J.S.D., Deravi, N.: Integrated person identification using voice and facial features. In: Proceedings of the IEE Colloquium on Image Processing for Security Application, pp. 4/1–4/5 (1997)
32. Choudhury, T., Reh, J.M., Pavlovic, V., Pentland, A.: Boosting and structure learning in dynamic Bayesian networks for audio-visual speaker detection. In: Proceedings of the IEEE International Conference on Pattern Recognition, vol. 3, pp. 789–794 (2002)
33. Cichocki, A., Zdunek, R., Amari, S.: Nonnegative matrix and tensor factorization. *IEEE Signal Process. Mag.* **25**(1), 142–145 (2008)
34. Compagnoni, M., Bestagini, P., Antonacci, F., Sarti, A., Tubaro, S.: Localization of acoustic sources through the fitting of propagation cones using multiple independent arrays. *IEEE Trans. Audio Speech Lang. Process.* **20**(7), 1964–1975 (2012)
35. Cover, T.M., Thomas, J.A.: *Elements of Information Theory*. Wiley, London (2006)
36. Cox, H., Zeskind, R., Kooij, T.: Practical supergain. *IEEE Trans. Acoust. Speech Signal Process.* **34**(3), 393–398 (1986)
37. Cristani, M., Bicego, M., Murino, V.: Audio-visual event recognition in surveillance video sequences. *IEEE Trans. Multimed.* **9**(2), 257–267 (2007)
38. Crocco, M., Bue, A.D., Murino, V.: A bilinear approach to the position self-calibration of multiple sensors. *IEEE Trans. Signal Process.* **60**(2), 660–673 (2012)
39. Cutler, R., Davis, L.: Look who’s talking: speaker detection using video and audio correlation. In: Proceedings of the IEEE International Conference on Multimedia & Expo, vol. 3, pp. 1589–1592. IEEE, New York (2000)
40. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893. IEEE, New York (2005)
41. D’Arca, E., Robertson, N., Hopgood, J.: Look who’s talking: Detecting the dominant speaker in a cluttered scenario. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (2014)
42. DiBiase, J., Silverman, H., Brandstein, M.: Robust localization in reverberant rooms. In: *Microphone Arrays*, pp. 157–180. Springer, New York (2001)
43. Dmochowski, J., Benesty, J., Affes, S.: A generalized steered response power method for computationally viable source localization. *IEEE Trans. Audio Speech Lang. Process.* **15**(8), 2510–2526 (2007)
44. Do, H., Silverman, H., Yu, Y.: A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 1121–1124. IEEE, New York (2007)
45. Doclo, S., Moonen, M.: GSVD-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Trans. Signal Process.* **50**(9), 2230–2244 (2002)
46. Duong, N.Q.K., Vincent, E., Gribonval, R.: Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Trans. Audio Speech Lang. Process.* **18**(7), 1830–1840 (2010)
47. Duong, N.Q.K., Vincent, E., Gribonval, R.: Spatial location priors for Gaussian model based reverberant audio source separation. *EURASIP J. Adv. Signal Process.* **2013**(1), 1–11 (2013)
48. Elko, G.W.: Spatial coherence functions for differential microphones in isotropic noise fields. In: *Microphone Arrays: Signal Processing Techniques and Applications*, pp. 61–85. Springer, New York (2001)
49. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition (2016). arXiv preprint arXiv:1604.06573
50. Févotte, C., Cardoso, J.F.: Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models. In: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 78–81 (2005)
51. Fisher, J., Darrell, T., Freeman, W.T., Viola, P., Fisher III, J.W.: Learning joint statistical models for audio-visual fusion and segregation. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 772–778 (2001)

52. FitzGerald, D., Cranitch, M., Coyle, E.: Extended nonnegative tensor factorisation models for musical sound source separation. *Comput. Intell. Neurosci.* **2008**, 15 pp. (2008). Article ID 872425; doi:[10.1155/2008/872425](https://doi.org/10.1155/2008/872425)
53. Fitzgerald, D., Cranitch, M., Coyle, E.: Using tensor factorisation models to separate drums from polyphonic music. In: *Proceedings of the International Conference on Digital Audio Effects (2009)*
54. Foucher, S., Lalibert, F., Boulianne, G., Gagnon, L.: A Dempster-Shafer based fusion approach for audio-visual speech recognition with application to large vocabulary French speech. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (2006)*
55. Frost, O.L.: An algorithm for linearly constrained adaptive array processing. *Proc. IEEE* **60**(8), 926–935 (1972)
56. Gandhi, A., Sharma, A., Biswas, A., Deshmukh, O.: Gethr-net: A generalized temporally hybrid recurrent neural network for multimodal information fusion (2016). arXiv preprint arXiv:1609.05281
57. Gehrig, T., Nickel, K., Ekenel, H., Klee, U., McDonough, J.: Kalman filters for audio-video source localization. In: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 118–121. IEEE, New York (2005)
58. Goecke, R., Millar, J.B.: Statistical analysis of the relationship between audio and video speech parameters for Australian English. In: *Proceedings of the ISCA Tutor Res Workshop Audit-Vis Speech Process*, pp. 133–138 (2003)
59. Gowdy, J.N., Subramanya, A., Bartels, C., Bilmes, J.A.: DBN based multi-stream models for audio-visual speech recognition. In: *Proceedings of the IEEE International Conference of Acoustics, Speech and Signal Processing (2004)*
60. Gravier, G., Potamianos, G., Neti, C.: Asynchrony modeling for audio-visual speech recognition. In: *Proceedings of the International Conference on Human Language Technology Research*, pp. 1–6. Morgan Kaufmann Publishers Inc., San Diego (2002)
61. Gribonval, R., Zibulevsky, M.: Sparse component analysis. In: *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, pp. 367–420. Academic, New York (2010)
62. Griffiths, L., Jim, C.: An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. Antennas Propag.* **30**(1), 27–34 (1982)
63. Gustafsson, T., Rao, B.D., Trivedi, M.: Source localization in reverberant environments: modeling and statistical analysis. *IEEE Trans. Speech Audio Process.* **11**, 791–803 (2003)
64. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. *Neural Comput.* **16**(12), 2639–2664 (2004)
65. Haykin, S.: *Adaptive Filter Theory*, 5th edn. Pearson Education, Upper Saddle River (2014)
66. Haykin, S., Justice, J.H., Owsley, N.L., Yen, J., Kak, A.C.: *Array Signal Processing*. Prentice-Hall, Inc., Englewood Cliffs (1985)
67. Hotelling, H.: Relations between two sets of variates. *Biometrika* **28**(3–4), 321–377 (1936)
68. Hu, D., Li, X., Lu, X.: Temporal multimodal learning in audiovisual speech recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)*
69. Huang, P.S., Zhuang, X., Hasegawa-Johnson, M.: Improving acoustic event detection using generalizable visual features and multi-modality modeling. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 349–352. IEEE, New York (2011)
70. Huang, Y., Benesty, J., Elko, G., Mersereau, R.: Real-time passive source localization: a practical linear-correction least-squares approach. *IEEE Trans. Speech Audio Process.* **9**(8), 943–956 (2001)
71. Ivanov, Y., Serre, T., Bouvrie, J.: Error weighted classifier combination for multi-modal human identification. Tech. Rep. MIT-CSAIL-TR-2005–081, MIT (2005)
72. Izadinia, H., Saleemi, I., Shah, M.: Multimodal analysis for identification and segmentation of moving-sounding objects. *IEEE Trans. Multimed.* **15**(2), 378–390 (2013)

73. Izumi, Y., Ono, N., Sagayama, S.: Sparseness-based 2CH BSS using the EM algorithm in reverberant environment. In: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 147–150 (2007)
74. Jaureguiberry, X., Vincent, E., Richard, G.: Fusion methods for speech enhancement and audio source separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(7), 1266–1279 (2016)
75. Jhuo, I.H., Ye, G., Gao, S., Liu, D., Jiang, Y.G., Lee, D., Chang, S.F.: Discovering joint audio-visual codewords for video event detection. *Mach. Vis. Appl.* **25**(1), 33–47 (2014)
76. Jiang, W., Loui, A.C.: Audio-visual grouplet: temporal audio-visual interactions for general video concept classification. In: Proceedings of the ACM International Conference on Multimedia, Scottsdale, pp. 123–132. (2011)
77. Jiang, Y.G., Zeng, X., Ye, G.: Columbia-UCF TRECVID2010 multimedia event detection: combining multiple modalities, contextual concepts, and temporal matching. In: Proceedings of the NIST TRECVID-2003 (2003)
78. Jiang, W., Cotton, C., Chang, S.F., Ellis, D., Loui, A.: Short-term audiovisual atoms for generic video concept classification. In: Proceedings of the ACM International Conference on Multimedia, pp. 5–14. ACM, New York (2009)
79. Jiang, Y.G., Bhattacharya, S., Chang, S.F., Shah, M.: High-level event recognition in unconstrained videos. *Int. J. Multimed. Inf. Retr.* **2**(2), 73–101 (2013)
80. Joder, C., Essid, S., Richard, G.: Temporal integration for audio classification with application to musical instrument classification. *IEEE Trans. Audio Speech Lang. Process.* **17**(1), 174–186 (2009). doi:[10.1109/TASL.2008.2007613](https://doi.org/10.1109/TASL.2008.2007613)
81. Jourjine, A., Rickard, S., Yilmaz, O.: Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2985–2988 (2000)
82. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725–1732 (2014)
83. Kay, J.: Feature discovery under contextual supervision using mutual information. In: Proceedings of the International Joint Conference on Neural Networks, vol. 4, pp. 79–84 (1992)
84. Kidron, E., Schechner, Y., Elad, M.: Pixels that sound. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 88–95 (2005)
85. Kijak, E., Gravier, G., Gros, P., Oisel, L., Bimbot, F.: HMM based structuring of tennis videos using visual and audio cues. In: Proceedings of the IEEE International Conference on Multimedia Expo, pp. 309–312. IEEE Computer Society, Washington (2003)
86. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(3), 226–239 (1998)
87. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. *SIAM Rev.* **51**(3), 455–500 (2009)
88. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
89. Kuhn, G.F.: Model for the interaural time differences in the azimuthal plane. *J. Acoust. Soc. Am.* **62**(1), 157–167 (1977)
90. Lai, P.L., Fyfe, C.: Kernel and nonlinear canonical correlation analysis. *Int. J. Neural Syst.* **10**(5), 365–378 (2000)
91. Levy, A., Gannot, S., Habets, E.: Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments. *IEEE Trans. Audio Speech Lang. Process.* **19**(6), 1540–1555 (2011)
92. Li, D., Dimitrova, N., Li, M., Sethi, I.: Multimedia content processing through cross-modal association. In: Proceedings of the ACM International Conference on Multimedia, Berkeley, CA (2003)

93. Lim, A., Nakamura, K., Nakadai, K., Ogata, T., Okuno, H.G.: Audio-visual musical instrument recognition. In: Proceedings of the National Convention Audio-V Information Processing Society (2011)
94. Liu, Q., Wang, W., Jackson, P.J., Barnard, M., Kittler, J., Chambers, J.: Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking. *IEEE Trans. Signal Process.* **61**(22), 5520–5535 (2013)
95. Liutkus, A., Durrieu, J.L., Daudet, L., Richard, G.: An overview of informed audio source separation. In: Proceedings of the International Workshop on Image Analysis for Multimedia Interactive Services, pp. 1–4. IEEE, New York (2013)
96. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)
97. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, vol. 249, p. 250 (2010)
98. Makino, S., Lee, T.W., Sawada, H.: Blind Speech Separation. Springer, New York (2007)
99. Mandel, M., Ellis, D.: EM localization and separation using interaural level and phase cues. In: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 275–278 (2007)
100. Mandel, M., Bressler, S., Shinn-Cunningham, B., Ellis, D.: Evaluating source separation algorithms with reverberant speech. *IEEE Trans. Audio Speech Lang. Process.* **18**(7), 1872–1883 (2010)
101. Maragos, P., Gros, P., Katsamanis, A., Papandreou, G.: Cross-modal integration for performance improving in multimedia: a review. In: *Multimodal Processing and Interaction*, pp. 1–46. Springer, New York (2008)
102. Marti, A., Cobos, M., Lopez, J., Escolano, J.: A steered response power iterative method for high-accuracy acoustic source localization. *J. Acoust. Soc. Am.* **134**(4), 2627–2630 (2013)
103. Metallinou, A., Lee, S., Narayanan, S.: Decision level combination of multiple modalities for recognition and analysis of emotional expression. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2462–2465 (2010)
104. Milani, S., Fontani, M., Bestagini, P., Barni, M., Piva, A., Tagliasacchi, M., Tubaro, S.: An overview on video forensics. *APSIPA Trans. Signal Inf. Process.* **1**, e2 (2012)
105. Monaci, G., Vanderghenst, P.: Audiovisual gestalts. In: Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, pp. 200–200 (2006)
106. Monaci, G., Jost, P., Vanderghenst, P., Mailhé, B., Lesage, S., Gribonval, R.: Learning multimodal dictionaries. *IEEE Trans. Image Process.* **16**(9), 2272–2283 (2007)
107. Monaci, G., Vanderghenst, P., Sommer, F.T.: Learning bimodal structure in audio-visual data. *IEEE Trans. Neural Netw.* **20**(12), 1898–1910 (2009)
108. Moore, B.C.J.: Introduction to the Psychology of Hearing. Macmillan, London (1977)
109. Murphy, K.P.: Dynamic Bayesian networks: representation, inference and learning. Ph.D. thesis, University of California, Berkeley (2002)
110. Naphade, M.R., Garg, A., Huang, T.S.: Audio-visual event detection using duration dependent input output markov models. In: Proceedings of the IEEE Workshop Content-Based Access Image and Video Libraries, pp. 39–43. IEEE, New York (2001)
111. Nefian, A.V., Liang, L., Pi, X., Xiaoxiang, L., Mao, C., Murphy, K.P.: A coupled HMM for audiovisual speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2. IEEE, New York (2002)
112. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the International Conference on Machine Learning, pp. 689–696 (2011)
113. Nguyen, V.T., Nguyen, D.L., Tran, M.T., Le, D.D., Duong, D.A., Satoh, S.: Query-adaptive late fusion with neural network for instance search. In: Proceedings of the IEEE International Workshop on Multimedia Signal Processing, pp. 1–6. IEEE, New York (2015)

114. Nikunen, J., Virtanen, T.: Direction of arrival based spatial covariance model for blind sound source separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(3), 727–739 (2014)
115. Omologo, M., Svaizer, P.: Acoustic event localization using a crosspower-spectrum phase based technique. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2. IEEE, New York (1994)
116. Otsuka, T., Ishiguro, K., Sawada, H., Okuno, H.G.: Bayesian nonparametrics for microphone array processing. *IEEE/ACM Trans. Audio Speech Lang. Proc.* **22**(2), 493–504 (2014)
117. Ozerov, A., Févotte, C.: Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **18**(3), 550–563 (2010)
118. Ozerov, A., Févotte, C., Blouet, R., Durrieu, J.L.: Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague (2011)
119. Ozerov, A., Vincent, E., Bimbot, F.: A general flexible framework for the handling of prior information in audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **20**(4), 1118–1133 (2012)
120. Parekh, S., Essid, S., Ozerov, A., Duong, N.Q.K., Pérez, P., Richard, G.: Motion informed audio source separation. In: *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, New Orleans (2017)
121. Parisi, R., Croene, P., Uncini, A.: Particle swarm localization of acoustic sources in the presence of reverberation. In: *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 4. IEEE, New York (2006)
122. Parra, L., Spence, C.: Convolutive blind separation of non-stationary sources. *IEEE Trans. Speech Audio Process.* **8**(3), 320–327 (2000)
123. Pertilä, P., Mieskolainen, M., Hämäläinen, M.: Closed-form self-localization of asynchronous microphone arrays. In: *Proceedings of the Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, pp. 139–144. IEEE, New York (2011)
124. Rocha, A., Scheirer, W., Boulton, T., Goldenstein, S.: Vision of the unseen: Current trends and challenges in digital image and video forensics. *ACM Comput. Surv.* **43**(4), 26 (2011)
125. Rokach, L.: Ensemble-based classifiers. *Artif. Intell. Rev.* **33**(1), 1–39 (2010). doi:[10.1007/s10462-009-9124-7](https://doi.org/10.1007/s10462-009-9124-7)
126. Roy, R., Kailath, T.: Esprit-estimation of signal parameters via rotational invariance techniques. *IEEE Trans. Acoust. Speech Signal Process.* **37**(7), 984–995 (1989)
127. Sadlier, D.A., O'Connor, N.E.: Event detection in field sports video using audio-visual features and a support vector machine. *IEEE Trans. Circuits Syst. Video Technol.* **15**(10), 1225–1233 (2005)
128. Sawada, H., Mukai, R., Araki, S., Makino, S.: A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Trans. Speech Audio Process.* **12**(5), 530–538 (2004)
129. Schau, H., Robinson, A.: Passive source localization employing intersecting spherical surfaces from time-of-arrival differences. *IEEE Trans. Acoust. Speech Signal Process.* **35**(8), 1223–1225 (1987)
130. Scheuing, J., Yang, B.: Disambiguation of tdoa estimation for multiple sources in reverberant environments. *IEEE Trans. Audio Speech Lang. Process.* **16**(8), 1479–1489 (2008)
131. Schmidt, R.: Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propag.* **34**(3), 276–280 (1986)
132. Sedighin, F., Babaie-Zadeh, M., Rivet, B., Jutten, C.: Two multimodal approaches for single microphone source separation. In: *Proceedings of the European Signal Processing Conference* (2016)
133. Seichepine, N., Essid, S., Févotte, C., Cappe, O.: Soft nonnegative matrix co-factorization with application to multimodal speaker diarization. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver (2013)
134. Seichepine, N., Essid, S., Févotte, C., Cappe, O.: Soft nonnegative matrix co-factorization. *IEEE Trans. Signal Process.* **PP**(99) (2014)

135. Serizel, R., Moonen, M., van Dijk, B., Wouters, J.: Low-rank approximation based multi-channel wiener filter algorithms for noise reduction with application in cochlear implants. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(4), 785–799 (2014)
136. Serizel, R., Bisot, V., Essid, S., Richard, G.: Machine listening techniques as a complement to video image analysis in forensics. In: *Proceedings of the IEEE International Conference on Image Processing*, pp. 948–952. IEEE, New York (2016)
137. Showen, R., Calhoun, R., Dunham, J.: Acoustic location of gunshots using combined angle of arrival and time of arrival measurements (2009). US Patent 7,474,589
138. Sigg, C., Fischer, B., Ommmer, B., Roth, V., Buhmann, J.: Nonnegative CCA for audiovisual source separation. In: *Proceedings of the IEEE Workshop Machine Learning and Signal Processing*, pp. 253–258. IEEE, New York (2007)
139. Smaragdis, P., Casey, M.: Audio visual independent components. In: *Proceedings of the International Symposium Independent Component Analysis and Blind Signal Separation*, pp. 709–714 (2003)
140. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep Boltzmann machines. In: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 2222–2230 (2012)
141. P. Stoica, Moses, R.: *Spectral Analysis of Signals*. Pearson Prentice Hall, Upper Saddle River, NJ (2005)
142. Strobel, N., Spors, S., Rabenstein, R.: Joint audio-video object localization and tracking. *IEEE Signal Process. Mag.* **18**(1), 22–31 (2001)
143. Tian, Y., Chen, Z., Yin, F.: Distributed Kalman filter-based speaker tracking in microphone array networks. *Appl. Acoust.* **89**, 71–77 (2015)
144. Togami, M., Hori, K.: Multichannel semi-blind source separation via local Gaussian modeling for acoustic echo reduction. In: *Proceedings of the European Signal Processing Conference* (2011)
145. Togami, M., Kawaguchi, Y.: Simultaneous optimization of acoustic echo reduction, speech dereverberation, and noise reduction against mutual interference. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(11), 1612–1623 (2014)
146. Trifa, V., Koene, A., Moren, J., Cheng, G.: Real-time acoustic source localization in noisy environments for human-robot multimodal interaction. In: *Proceedings of the IEEE International Symposium on Robots and Human Interactive Communication* (2007)
147. Valente, S., Tagliasacchi, M., Antonacci, F., Bestagini, P., Sarti, A., Tubaro, S.: Geometric calibration of distributed microphone arrays from acoustic source correspondences. In: *Proceedings of the IEEE International Workshop on Multimedia Signal Processing*, pp. 13–18 (2010)
148. Valin, J., Michaud, F., Rouat, J.: Robust 3d localization and tracking of sound sources using beamforming and particle filtering. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4. IEEE, New York (2006)
149. Velivelli, A., Ngo, C.W., Huang, T.S.: Detection of documentary scene changes by audio-visual fusion. In: *Proceedings of the International Conference on Image and Video Retrieval*, pp. 227–238. Springer, New York (2003)
150. Vincent, E., Bertin, N., Gribonval, R., Bimbot, F.: From blind to guided audio source separation: how models and side information can improve the separation of sound. *IEEE Signal Process. Mag.* **31**(3), 107–115 (2014)
151. Vuegen, L., Broeck, B.V.D., Karsmakers, P., hamme, H.V., Vanrumste, B.: Automatic monitoring of activities of daily living based on real-life acoustic sensor data: a preliminary study. In: *Proceedings of the International Workshop on Speech and Language Processing for Assistive Technologies*, pp. 113–118 (2013)
152. Wang, D.L.: Time-frequency masking for speech separation and its potential for hearing aid design. *Trends Amplif.* **12**(4), 332–352 (2008)
153. Wang, H., Chu, P.: Voice source localization for automatic camera pointing system in videoconferencing. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (1997)

154. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. *Int. J. Comput. Vis.* **103**(1), 60–79 (2013)
155. Ward, D.B., Lehmann, E.A., Williamson, R.C.: Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Trans. Speech Audio Process.* **11**(6), 826–836 (2003)
156. Wilkins, P., Adamek, T., Byrne, D., Jones, G., Lee, H., Keenan, G., Mcguinness, K., O'Connor, N.E., Smeaton, A.F., Amin, A., Obrenovic, Z., Benmokhtar, R., Galmar, E., Huet, B., Essid, S., Landais, R., Vallet, F., Papadopoulos, G.T., Vrochidis, S., Mezaris, V., Kompatsiaris, I., Spyrou, E., Avrithis, Y., Morzinger, R., Schallauer, P., Bailer, W., Piatrik, T., Chandramouli, K., Izquierdo, E., Haller, M., Goldmann, L., Samour, A., Cobet, A., Sikora, T., Praks, P.: K-space at TRECVID 2007. In: *TRECVID 2007* (2007)
157. Wu, Y., Lin, C.Y.Y., Chang, E.Y., Smith, J.R.: Multimodal information fusion for video concept detection. In: *Proceedings of the IEEE International Conference on Image Processing*, vol. 4, pp. 2391–2394. IEEE, Singapore (2004)
158. Wu, Z., Jiang, Y.G., Wang, J., Pu, J., Xue, X.: Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In: *Proceedings of the ACM International Conference on Multimedia*, pp. 167–176. ACM, New York (2014)
159. Yilmaz, K., Cemgil, A.T.: Probabilistic latent tensor factorisation. In: *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation*, pp. 346–353 (2010)
160. Yokoya, N., Yairi, T., Iwasaki, A.: Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Trans. Geosci. Remote Sens.* **50**(2), 528–537 (2012)
161. Yoo, J., Choi, S.: Matrix co-factorization on compressed sensing. In: *Proceedings of the International Joint Conference on Artificial Intelligence* (2011)
162. Yost, W.A.: Discriminations of interaural phase differences. *J. Acoust. Soc. Am.* **55**(6), 1299–1303 (1974)
163. Yuhas, B.P., Goldstein, M.H., Sejnowski, T.J.: Integration of acoustic and visual speech signals using neural networks. *IEEE Commun. Mag.* **27**(11), 65–71 (1989)
164. Zhang, Q., Chen, Z., Yin, F.: Distributed marginalized auxiliary particle filter for speaker tracking in distributed microphone networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(11), 1921–1934 (2016)
165. Zotkin, D.N., Duraiswami, R.: Accelerated speech source localization via a hierarchical search of steered response power. *IEEE Trans. Speech Audio Process.* **12**(5), 499–508 (2004)