Multimedia Systems (2010) 16:345–379
DOI 10.1007/s00530-010-0182-0

REGULAR PAPER

Multimodal fusion for multimedia analysis: a survey

Pradeep K. Atrey · M. Anwar Hossain · Abdulmotaleb El Saddik · Mohan S. Kankanhalli

Received: 8 January 2009 / Accepted: 9 March 2010 / Published online: 4 April 2010

Abstract This survey aims at providing multimedia researchers with a state-of-the-art overview of fusion strategies, which are used for combining multiple modalities in order to accomplish various multimedia analysis tasks. The existing literature on multimodal fusion research is presented through several classifications based on the fusion methodology and the level of fusion (feature, decision, and hybrid). The fusion methods are described from the perspective of the basic concept, advantages, weaknesses, and their usage in various analysis tasks as reported in the literature. Moreover, several distinctive issues that influence a multimodal fusion process such as, the use of correlation and independence, confidence level, contextual information, synchronization between different modalities, and the optimal modality selection are also highlighted. Finally, we present the open issues for further research in the area of multimodal fusion.

Communicated by Wu-chi Feng.

P. K. Atrey (✉)
Department of Applied Computer Science,
University of Winnipeg, Winnipeg, Canada
e-mail: p.atrey@uwinnipeg.ca

M. A. Hossain · A. El Saddik
Multimedia Communications Research Laboratory,
University of Ottawa, Ottawa, Canada
e-mail: anwar@mcrlab.uottawa.ca

A. El Saddik
e-mail: abed@mcrlab.uottawa.ca

M. S. Kankanhalli
School of Computing, National University of Singapore,
Singapore, Singapore
e-mail: mohan@comp.nus.edu.sg

Keywords Multimodal information fusion · Multimedia analysis

1 Introduction

In recent times, multimodal fusion has gained much attention of many researchers due to the benefit it provides for various multimedia analysis tasks. The integration of multiple media, their associated features, or the intermediate decisions in order to perform an analysis task is referred to as multimodal fusion. A multimedia analysis task involves processing of multimodal data in order to obtain valuable insights about the data, a situation, or a higher level activity. Examples of multimedia analysis tasks include semantic concept detection, audio-visual speaker detection, human tracking, event detection, etc. Multimedia data used for these tasks could be sensory (such as audio, video, RFID) as well as non-sensory (such as WWW resources, database). These media and related features are fused together for the accomplishment of various analysis tasks. The fusion of multiple modalities can provide complementary information and increase the accuracy of the overall decision making process. For example, fusion of audio-visual features along with other textual information have become more effective in detecting events from a team sports video [149], which would otherwise not be possible by using a single medium.

The benefit of multimodal fusion comes with a certain cost and complexity in the analysis process. This is due to the different characteristics of the involved modalities, which are briefly stated in the following:

- Different media are usually captured in different formats and at different rates. For example, a video

may be captured at a frame rate that could be different from the rate at which audio samples are obtained, or even two video sources could have different frame rates. Therefore, the fusion process needs to address this asynchrony to better accomplish a task.

- The processing time of different types of media streams are dissimilar, which influences the fusion strategy that needs to be adopted.
- The modalities may be correlated or independent. The correlation can be perceived at different levels, such as the correlation among low-level features that are extracted from different media streams and the correlation among semantic-level decisions that are obtained based on different streams. On the other hand, the independence among the modalities is also important as it may provide additional cues in obtaining a decision. When fusing multiple modalities, this correlation and independence may equally provide valuable insight based on a particular scenario or context.
- The different modalities usually have varying confidence levels in accomplishing different tasks. For example, for detecting the event of a human crying, we may have higher confidence in an audio modality than a video modality.
- The capturing and processing of media streams may involve certain costs, which may influence the fusion process. The cost may be incurred in units of time, money or other units of measure. For instance, the task of object localization could be accomplished cheaply by using a RFID sensor compared to using a video camera.

The above characteristics of multiple modalities influence the way the fusion process is carried out. Due to these varying characteristics and the objective tasks that need to be carried out, several challenges may appear in the multimodal fusion process as stated in the following:

- *Levels of fusion*. One of the earliest considerations is to decide what strategy to follow when fusing multiple modalities. The most widely used strategy is to fuse the information at the feature level, which is also known as early fusion. The other approach is decision level fusion or late fusion [45, 121] which fuses multiple modalities in the semantic space. A combination of these approaches is also practiced as the hybrid fusion approach [144].
- *How to fuse?* There are several methods that are used in fusing different modalities. These methods are particularly suitable under different settings and are described in this paper in greater detail. The discussion also includes how the fusion process utilizes the feature and decision level correlation among the modalities [103], and how the contextual [100] and the confidence information [18] influences the overall fusion process.
- *When to fuse?* The time when the fusion should take place is an important consideration in the multimodal fusion process. Certain characteristics of media, such as varying data capture rates and processing time of the media, poses challenges on how to synchronize the overall process of fusion. Often this has been addressed by performing the multimedia analysis tasks (such as event detection) over a timeline [29]. A timeline refers to a measurable span of time with information denoted at designated points. The timeline-based accomplishment of a task requires identification of designated points at which fusion of data or information should take place. Due to the asynchrony and diversity among streams and due to the fact that different analysis tasks are performed at different granularity levels in time, the identification of these designated points, i.e. when the fusion should take place, is a challenging issue [8].
- *What to fuse?* The different modalities used in a fusion process may provide complementary or contradictory information and therefore knowing which modalities are contributing towards accomplishing an analysis task needs to be understood. This is also related to finding the optimal number of media streams [9, 143] or feature sets required to accomplish an analysis task under the specified constraints. If the most suitable subset is unavailable, can one use alternate streams without much loss of cost-effectiveness and confidence?

This paper presents a survey of the research related to multimodal fusion for multimedia analysis in light of the above challenges. Existing surveys in this direction are mostly focused on a particular aspect of the analysis task, such as multimodal video indexing [26, 120]; automatic audio-visual speech recognition [106]; biometric audio-visual speech synchrony [20]; multi-sensor management for information fusion [146]; face recognition [153]; multimodal human computer interaction [60, 97]; audio-visual biometric [5]; multi-sensor fusion [79] and many others. In spite of these literatures, a comprehensive survey focusing on the different methodologies and issues related to multimodal fusion for performing different multimedia analysis tasks is still missing. The presented survey aims to contribute in this direction. The fusion problems have also been addressed in other domains such as machine learning [48], data mining [24] and information retrieval [133], however, the focus of this paper is restricted to the multimedia research domain.

Consequently, this work comments on the state-of-the-art literature that uses different multimodal fusion strategies for various analysis tasks such as audio-visual person

tracking, video summarization, multimodal dialog understanding, speech recognition and so forth. It also presents several classifications of the existing literature based on the fusion methodology and the level of fusion. Various issues such as the use of correlation, context and confidence, and the optimal modality selection that influences the performance of a multimodal fusion process is also critically discussed.

The remainder of this paper is organized as follows. In Sect. 2, we first address the issue *levels of fusion* and accordingly describe three levels (feature, decision and hybrid) of multimodal fusion, their characteristics, advantages and limitations. Section 3 addresses the issue *how to fuse* by describing the various fusion methods that have been used for multimedia analysis. These fusion methods have been elaborated under three different categories—the rule-based methods, the estimation-based methods, and the classification-based methods. In this section, we analyze various related works from the perspective of the level of fusion, the modality used, and the multimedia analysis task performed. A discussion regarding the different fusion methodologies and the works we analyzed is also presented here. Some other issues (e.g. the use of correlation, confidence, and the context), also related to *how to fuse*, are described in Sect. 4. This section further elaborates the issues *when to fuse* (the synchronization), and *what to fuse* (the optimal media selection). Section 5 provides a brief overview of the publicly available data sets and evaluation measures in multimodal fusion research. Finally, Sect. 6 concludes the paper by pointing out the open issues and possible avenues of further research in the area of multimodal fusion for multimedia analysis.

2 Levels of fusion

The fusion of different modalities is generally performed at two levels: *feature level* or *early fusion* and *decision level* or *late fusion* [3, 45, 121]. Some researchers have also followed a hybrid approach by performing fusion at the feature as well as the decision level.

Figure 1 shows different variants of the feature, decision, and hybrid level fusion strategies. We now describe the three levels of fusion and highlight their pros and cons. Various works that have adopted different fusion models at different levels (feature, decision and hybrid) in different scenarios will be discussed in Sect. 3.

2.1 Feature level multimodal fusion

In the feature level or early fusion approach, the features extracted from input data are first combined and then sent as input to a single analysis unit (AU) that performs the analysis task. Here, features refer to some distinguishable properties of a media stream. For example, the feature fusion (FF) unit merges the multimodal features such as skin color and motion cues into a larger feature vector which is taken as the input to the face detection unit in order to detect a face. An illustration of this is provided in Fig. 1. While Fig. 1a shows an AU that receives a set of either features or decisions and provides a semantic-level decision, Fig. 1b shows a FF unit that receives a set of features F_1 to F_n and combines them into a feature vector $F_{1,n}$. Figure 1d shows an instance of the feature level multimodal analysis task in which the extracted features are first fused using a FF unit and then the combined feature vector is passed to an AU for analysis.

In the feature level fusion approach, the number of features extracted from different modalities may be numerous, which may be summarized as [138, 150]:

- *Visual features*. It may include features based on color (e.g. color histogram), texture (e.g. measures of coarseness, directionality, contrast), shape (e.g. blobs), and so on. These features are extracted from the entire image, fixed-sized patches or blocks, segmented image blobs or automatically detected feature points.
- *Text features*. The textual features can be extracted from the automatic speech recognizer (ASR) transcript, video optical character recognition (OCR), video closed caption text, and production metadata.
- *Audio features*. The audio features may be generated based on the short time Fourier transform including the fast Fourier transform (FFT), mel-frequency cepstral coefficient (MFCC) together with other features such as zero crossing rate (ZCR), linear predictive coding (LPC), volume standard deviation, non-silence ratio, spectral centroid and pitch.
- *Motion features*. This can be represented in the form of kinetic energy which measures the pixel variation within a shot, motion direction and magnitude histogram, optical flows and motion patterns in specific directions.
- *Metadata*. The metadata features are used as supplementary information in the production process, such as the name, the time stamp, the source of an image or video as well as the duration and location of shots. They can provide extra information to text or visual features.

The feature level fusion is advantageous in that it can utilize the correlation between multiple features from different modalities at an early stage which helps in better task accomplishment. Also, it requires only one learning phase on the combined feature vector [121]. However, in this approach it is hard to represent the time synchronization between the multimodal features [144]. This is because

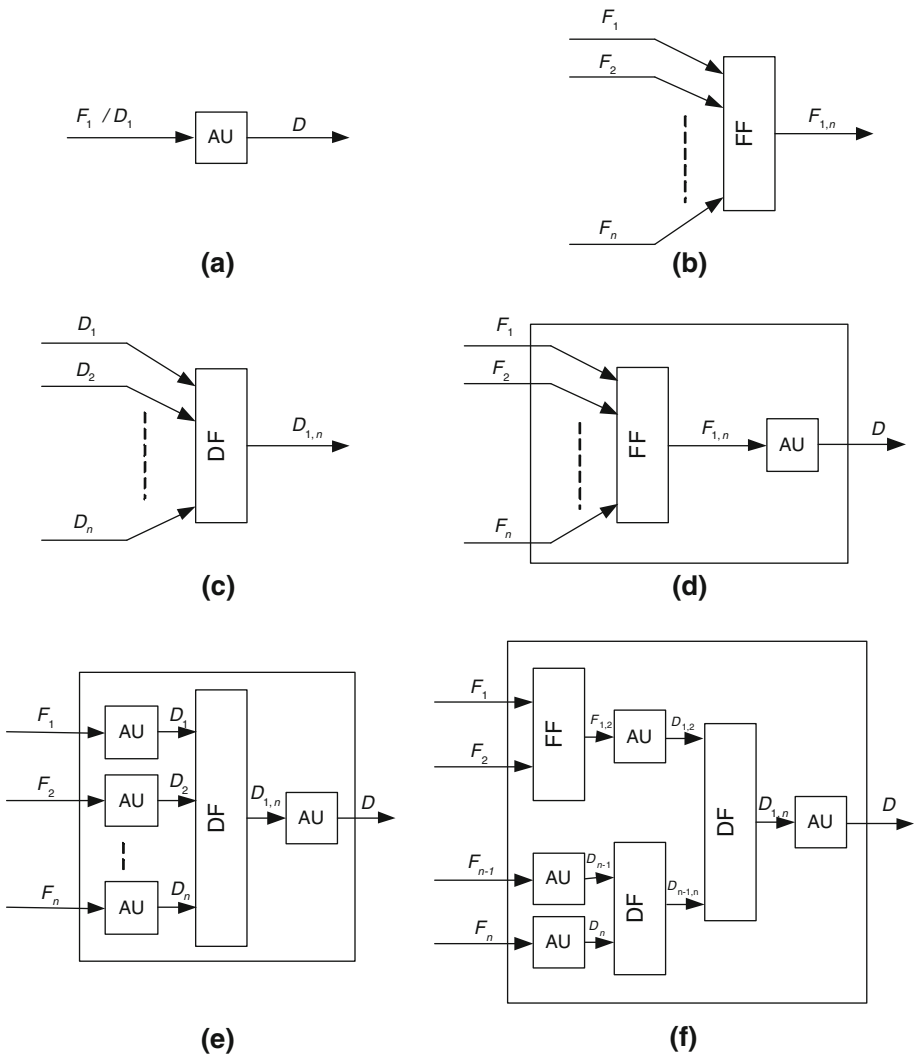

Fig. 1 Multimodal fusion strategies and conventions as used in this paper. **a** Analysis unit, **b** feature fusion unit, **c** decision fusion unit, **d** feature level multimodal analysis, **e** decision level multimodal analysis, **f** hybrid multimodal analysis

the features from different but closely coupled modalities could be extracted at different times. Moreover, the features to be fused should be represented in the same format before fusion. In addition, the increase in the number of modalities makes it difficult to learn the cross-correlation among the heterogeneous features. Various approaches to resolve the synchronization problem are discussed in Sect. 4.2.

Several researchers have adopted the early fusion approach for different multimedia analysis tasks. For instance, Nefian et al. [86] have adopted an early fusion approach in combining audio and visual features for speech recognition.

2.2 Decision level multimodal fusion

In the decision level or late fusion approach, the analysis units first provide the local decisions D_1 to D_n (see Fig. 1) that are obtained based on individual features F_1 to F_n. The local decisions are then combined using a decision fusion (DF) unit to make a fused decision vector that is analyzed further to obtain a final decision D about the task or the hypothesis. Here, a decision is the output of an analysis unit at the semantic level. An illustration of DF unit is provided in Fig. 1c whereas Fig. 1e shows an instance of the decision level multimodal analysis in which the decisions obtained from various AUs are fused using a DF unit and the combined decision vector is further processed by an AU.

The decision level fusion strategy has many advantages over feature fusion. For instance, unlike feature level fusion, where the features from different modalities (e.g. audio and video) may have different representations, the decisions (at the semantic level) usually have the same representation. Therefore, the fusion of decisions becomes easier. Moreover, the decision level fusion strategy offers scalability (i.e. graceful upgradation or degradation) in terms of the modalities used in the fusion process, which is difficult to achieve in the feature level fusion [9]. Another advantage of late fusion strategy is that it allows us to use

the most suitable methods for analyzing each single modality, such as hidden Markov model (HMM) for audio and support vector machine (SVM) for image. This provides much more flexibility than the early fusion.

On the other hand, the disadvantage of the late fusion approach lies in its failure to utilize the feature level correlation among modalities. Moreover, as different classifiers are used to obtain the local decisions, the learning process for them becomes tedious and time-consuming.

Several researchers have successfully adopted the decision level fusion strategy. For example, Iyenger et al. [57] performed fusion of decisions obtained from a face detector and a speech recognizer along with their synchrony score by adopting two approaches—a linear weighted sum and a linear weighted product.

2.3 Hybrid multimodal fusion

To exploit the advantages of both the feature level and the decision level fusion strategies, several researchers have opted to use a hybrid fusion strategy, which is a combination of both feature and decision level strategies. An illustration of the hybrid level strategy is presented in Fig. 1f where the features are first fused by a FF unit and then the feature vector is analyzed by an AU. At the same time, other individual features are analyzed by different AUs and their decisions are fused using a DF unit. Finally, all the decisions obtained from the previous stages are further fused by a DF to obtain the final decision.

A hybrid fusion approach can utilize the advantages of both early and late fusion strategies. Therefore, many researchers ([16, 88, 149], etc.) have used the hybrid fusion strategy to solve various kinds of multimedia analysis problems.

3 Methods for multimodal fusion

In this section, we provide an overview of the different fusion methods that have been used by the multimedia researchers to perform various multimedia analysis tasks. The advantages and the drawbacks of each method are also highlighted. The fusion methods are divided into the following three categories: rule-based methods, classification-based methods, and estimation-based methods (as shown in Fig. 2). This categorization is based on the basic nature of these methods and it inherently means the classification of the problem space, such as, a problem of estimating parameters is solved by estimation-based methods. Similarly the problem of obtaining a decision based on certain observation can be solved by classification-based or rule-based methods. However, if the observation is obtained from different modalities, the method would require fusion

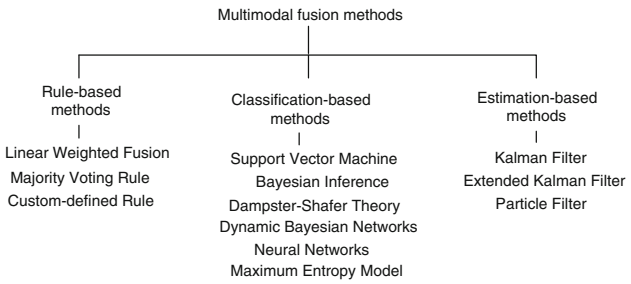

Fig. 2 A categorization of the fusion methods

of the observation scores before estimation or making a classification decision.

While the next three sections (Sect. 3.1–3.3) have been devoted to the above three classes of fusion methods; in the last section (Sect. 3.4), we present a comparative analysis of all the fusion methods.

3.1 Rule-based fusion methods

The rule-based fusion method includes a variety of basic rules of combining multimodal information. These include statistical rule-based methods such as linear weighted fusion (sum and product), MAX, MIN, AND, OR, majority voting. The work by Kittler et al. [69] has provided the theoretical introduction of these rules. In addition to these rules, there are custom-defined rules that are constructed for the specific application perspective. The rule-based schemes generally perform well if the quality of temporal alignment between different modalities is good. In the following, we describe some representative works that have adopted the rule-based fusion strategy.

3.1.1 Linear weighted fusion

Linear weighted fusion is one of the simplest and most widely used methods. In this method, the information obtained from different modalities is combined in a linear fashion. The information could be the low-level features (e.g. color and motion cues in video frames) [136] or the semantic-level decisions (i.e. occurrence of an event) [90]. To combine the information, one may assign normalized weights to different modalities. In literature, there are various methods for weight normalization such as min–max, decimal scaling, z score, tan$h$-estimators and sigmoid function [61]. Each of these methods have pros and cons. The min–max, decimal scaling and z score methods are preferred when the matching scores (minimum and maximum values for min–max, maximum for decimal scaling and mean and standard deviation for z score) of the individual modalities can be easily computed. But these methods are sensitive to outliers. On the other hand, tan$h$

normalization method is both robust and efficient but requires estimation of the parameters using training. Note that the absence of prior knowledge of the weights usually equals the weight assigned to them.

The general methodology of linear fusion can be described as follows. Let I_i, $1 \leq i \leq n$ be a feature vector obtained from ith media source (e.g. audio, video etc.) or a decision obtained from a classifier.[1] Also, let w_i, $1 \leq i \leq n$ be the normalized weight assigned to the ith media source or classifier. These vectors, assuming that they have the same dimensions, are combined by using sum or product operators and used by the classifiers to provide a high-level decision. This is shown in Eqs. 1 and 2, which are as follows:

$$I = \sum_{i=1}^{n} w_i \times I_i \tag{1}$$

$$I = \prod_{i=1}^{n} I_i^{w_i} \tag{2}$$

This method is computationally less expensive compared to other methods. However, a fusion system needs to determine and adjust the weights for the optimal accomplishment of a task.

Several researchers have adopted the linear fusion strategy at the feature level for performing various multimedia analysis tasks. Examples include Foresti and Snidaro [40], Yang et al. [152] for detecting and tracking people, and Wang et al. [136] and Kankanhalli et al. [67] for video surveillance and traffic monitoring. The linear fusion strategy has also been adopted at the decision level by several researchers. These include Neti et al. [87] for speaker recognition and speech event detection, Iyengar et al. [57] for monologue detection, Iyengar et al. [58] for semantic concept detection and annotation in video, Lucey et al. [78] for spoken word recognition, Hua and Zhang [55] for image retrieval, McDonald and Smeaton [83] for video shot retrieval and Jaffre and Pinquier [59] for person identification. We briefly describe these works in the following.

Foresti and Snidaro [40] used a linear weighted sum method to fuse trajectory information of the objects. The video data from each sensor in a distributed sensor network is processed for moving object detection (e.g. a blob). Once the blob locations are extracted from all sensors, their trajectory coordinates are averaged in a linear weighted fashion in order to estimate the correct location of the blob. The authors have also assigned weights to different sensors; however, the determination of these weights has been left to the user. Similar to [40], Yang et al. [152] also performed linear weighted fusion of the location information of the objects. However, unlike Foresti and Snidaro [40], Yang et al. [152] assigned equal weights to the different modalities.

The linear weighted sum strategy at the feature level has also been proposed by Wang et al. [136] for human tracking. In this work, the authors have fused several spatial cues such as color, motion and texture by assigning appropriate weights to them. However, in the fusion process, the issue of how different weights should be assigned to different cues has not been discussed. This work was extended by Kankanhalli et al. [67] for face detection, monologue detection, and traffic monitoring. In both works, the authors used a sigmoid function to normalize the weights of different modalities.

Neti et al. [87] obtained individual decisions for speaker recognition and speech event detection from audio features (e.g. phonemes) and visual features (e.g. visemes). They adopted a linear weighted sum strategy to fuse these individual decisions. The authors used the training data to determine the relative reliability of the different modalities and accordingly adjusted their weights. Similar to this fusion approach, Iyengar et al. [57] fused multiple modalities (face, speech and the synchrony score between them) by adopting two approaches at the decision level—a linear weighted sum and a linear weighted product. This methodology was applied for monologue detection. The synchrony or correlation between face and speech has been computed in terms of mutual information between them by considering the audio and video features as locally Gaussian distributed. The mutual information is a measure of the information of one modality conveyed about another. The weights of the different modalities have been determined at the training stage. While fusing different modalities, the authors have found the linear weighted sum approach to be a better option than the linear weighted product for their data set. This approach was later extended for semantic concept detection and annotation in video by Iyengar et al. [58]. Similar to [57], the linear weighted product fusion strategy has also been adopted by Jaffre and Pinquier [59] for fusing different modalities. In this work, the authors have proposed a multimodal person identification system by automatically associating voice and image using a standard product rule. The association is done through fusion of video and audio indexes. The proposed work used a common indexing mechanism for both audio and video based on frame-by-frame analysis. The audio and video indexes were fused using a product fusion rule at the late stage.

In another work, Lucey et al. [78] performed a linear weighted fusion for the recognition of spoken words. The word recognizer modules, which work on audio and video data separately, provided decisions about a word in terms

[1] To maintain consistency, we will use these notations for modalities in rest of this paper.

of the log likelihoods. These decisions are linearly fused by assigning weights to them. To determine the weights of the two decision components, the authors have chosen the discrete values (0, 0.5 and 1), which is a simple but non-realistic choice.

A decision level fusion scheme proposed by Hua and Zhang [55] is based on the human's psychological observations which they call "attention". The core idea of this approach is to fuse the decisions taken based on different cues such as the strength of a sound, the speed of a motion, the size of an object and so forth. These cues are considered as the attention properties and are measured by obtaining the set of features including color histogram, color moment, wavelet, block wavelet, correlogram, and blocked correlogram. The authors proposed a new fusion function which they call "attention fusion function". This new function is a variation of a linear weighted sum strategy and is derived by adding the difference of two decisions to their average (please refer to [55] for formalism). The authors have demonstrated the utility of the proposed attention based fusion model for image retrieval. Experimental results have shown that the proposed approach performed better in comparison to average or maximal fusion rules.

In the context of video retrieval, McDonald and Smeaton [83] have employed a decision level linear weighted fusion strategy to combine the normalized scores and ranks of the retrieval results. The normalization was performed using max–min method. The video shots were retrieved using different modalities such as text and multiple visual features (color, edge and texture). In this work, the authors found that the combining of scores with different weights has been best for combining text and visual results for TRECVid type searches, while combining scores and ranks with equal weights have been best for combining multiple features for a single query image. A similar approach was adopted by Yan et al. [151] for re-ranking the video. In this work, the authors used a linear weighted fusion strategy at the decision level in order to combine the retrieval scores obtained based on text and other modalities such as audio, video and motion.

From the works discussed above, it is observed that the optimal weight assignment is the major drawback of the linear weighted fusion method. The issue of finding the appropriate weight (or confidence level) for different modalities is an open research issue. This issue is further elaborated in Sect. 4.1.2.

3.1.2 Majority voting

Majority voting is a special case of weighted combination with all weights to be equal. In majority voting based fusion, the final decision is the one where the majority of the classifiers reach a similar decision [113]. For example, Radova and Psutka [108] have presented a speaker identification system by employing multiple classifiers. Here, the raw speech samples from the speaker are treated as features. From the speech samples, a set of patterns are identified for each speaker. The pattern usually contains a current utterance of several vowels. Each pattern is classified by two different classifiers. The output scores of all the classifiers were fused in a late integration approach to obtain the majority decision regarding the identity of the unknown speaker.

3.1.3 Custom-defined rules

Unlike the above approaches that use standard statistical rules, Pfleger [100] presented a production rule-based decision level fusion approach for integrating inputs from pen and speech modality. In this approach, each input modality (e.g. pen input) is interpreted within its context of use, which is determined based on the previously recognized input events and dialog states belonging to the same user turn. The production rule consists of a weighting factor and a condition-action part. These rules are further divided into three classes that work together to contribute to the fusion process. First, the synchronization rules are applied to track the processing state of the individual recognizer (e.g. speech recognizer) and in case of pending recognition results the other classes of rules are not fired to ensure synchronization. Second, the rules for multimodal event interpretation are used to determine which of the input events has the lead and need to be integrated. Furthermore, there may be conflicting events due to the recognition or interpretation error, which are addresses by obtaining the event with highest score. Third, the rules for unimodal interpretations are adopted when one of the recognizers do not produce any meaningful result, for example a time-out by one recognizer, which will lead to a single modality based decision making. This approach is further extended [101] and applied for discourse processing in a multiparty dialog scenario.

In another work, Holzapfel et al. [49] showed an example of multimodal integration approach using custom-defined rules. The authors combined speech and 3D pointing gestures as a means of natural interaction with a robot in a kitchen. Multimodal fusion is performed at the decision level based on the n-best lists generated by each of the event parsers. Their experiments showed that there is a close correlation in time of speech and gesture. Similarly, in [32], a rule-based system has been proposed for fusion of speech and 2D gestures in human computer interaction. Here the audio and gesture modalities are fused at the decision level. A drawback of these approaches is the

overhead to determine the best action based on *n*-best fused input.

In addition to the video, audio and gesture, other modalities such as closed caption text and external metadata have been used for several applications such as video indexing and content analysis for team sports videos. On this account, Babaguchi et al. [12] presented a knowledge-based technique to leverage the closed caption text of broadcast video streams for indexing video shots based on the temporal correspondence between them. The closed caption text features are extracted as keywords and the video features are extracted as temporal changes of color distribution. This work presumably integrates textual and visual modalities using a late fusion strategy.

3.1.4 Remarks on rule-based fusion methods

A summary of all the works (related to the rule-based fusion methods) described above is provided in Table 1. As can be seen from the table, in rule-based fusion category, linear weighted fusion method has been widely used by researchers. It is a simple as well as computationally less expensive approach. This method performs well if the weights of different modalities are appropriately determined, which has been a major issue in using this method. In the existing literature, this method has been used for face detection, human tracking, monologue detection, speech and speaker recognition, image and video retrieval, and person identification. On the other hand, the fusion using custom-defined rules has the flexibility of adding rules based on the requirements. However, in general, these rules are domain specific and defining the rules requires proper knowledge of the domain. This fusion method is widely used in the domain of multimodal dialog systems and sports video analysis.

3.2 Classification-based fusion methods

This category of methods includes a range of classification techniques that have been used to classify the multimodal observation into one of the pre-defined classes. The methods in this category are the support vector machine,

Table 1 A list of the representative works in the rule-based fusion methods category

Fusion method	Level of fusion	The work	Modalities	Multimedia analysis task
Linear weighted fusion	Feature	Foresti and Snidaro [40]	Video (trajectory coordinates)	Human tracking
		Wang et al. [136]	Video (color, motion and texture	Human tracking
		Yang et al. [152]	Video (trajectory coordinates)	Human tracking
		Kankanhalli et al. [67]	Video (color, motion and texture)	Face detection, monologue detection and traffic monitoring
	Decision	Neti et al. [87]	Audio (phonemes) and visual (visemes)	Speaker recognition
		Lucey et al. [78]	Audio (MFCC), video (Eigenlip)	Spoken word recognition
		Iyenger et al. [57, 58]	Audio (MFCC), video (DCT of the face region) and the synchrony score	Monologue detection, semantic concept detection and annotation in video
		Hua and Zhang [55]	Image (six features: color histogram, color moment, wavelet, block wavelet, correlagram, blocked correlagram)	Image retrieval
		Yan et al. [151]	Text (closed caption, video OCR), audio, video (color, edge and texture histogram), motion	Video retrieval
		McDonald and Smeaton [83]	Text and video (color, edge and texture)	Video retrieval
		Jaffre and Pinquier [59]	Audio, video index	Person identification from audio-visual sources
Majority voting rule	Decision	Radova and Psutka [108]	Raw speech (set of patterns)	Speaker identification from audio sources
Custom-defined rules	Decision	Babaguchi et al. [12]	Visual (color), closed caption text (keywords)	Semantic sports video indexing
		Corradini et al. [32]	Speech, 2D gesture	Human computer interaction
		Holzapfel et al. [49]	Speech, 3D pointing gesture	Multimodal interaction with robot
		Pfleger [100]	Pen gesture, speech	Multimodal dialog system

Bayesian inference, Dempster–Shafer theory, dynamic Bayesian networks, neural networks and maximum entropy model. Note that we can further classify these methods as generative and discriminative models from the machine learning perspective. For example, Bayesian inference and dynamic Bayesian networks are generative models, while support vector machine and neural networks are discriminative models. However, we skip further discussion on such classification for brevity.

3.2.1 Support vector machine

Support vector machine (SVM) [23] has become increasingly popular for data classification and related tasks. More specifically, in the domain of multimedia, SVMs are being used for different tasks including feature categorization, concept classification, face detection, text categorization, modality fusion, etc. Basically SVM is considered as a supervised learning method and is used as an optimal binary linear classifier, where a set of input data vectors are partitioned as belonging to either one of the two learned classes. From the perspective of multimodal fusion, SVM is used to solve a pattern classification problem, where the input to this classifier is the scores given by the individual classifier. The basic SVM method is extended to create a non-linear classifier by using the kernel concept, where every dot product in the basic SVM formalism is replaced using a non-linear kernel function.

Many existing literature use the SVM-based fusion scheme. Adams et al. [3] adopted a late fusion approach in order to detect semantic concepts (e.g. sky, fire-smoke) in videos using visual, audio and textual modalities. They use a discriminate learning approach while fusing different modalities at the semantic level. For example, the scores of all intermediate concept classifiers are used to construct a vector that is passed as the semantic feature in SVM as shown in Fig. 3. This figure depicts that audio, video and text scores are combined in a high-dimensional vector before being classified by SVM. The black and white dots in the figure represent two semantic concepts. A similar approach has been adopted by Iyengar et al. [58] for concept detection and annotation in video.

Wu et al. [141] reported two approaches to study the optimal combination of multimodal information for video concept detection, which are gradient-descent-optimization linear fusion (GLF) and the super-kernel nonlinear fusion (NLF). In GLF, an individual kernel matrix is first constructed for each modality providing a partial view of the target concept. The individual kernel matrices are then fused based on a weighted linear combination scheme. Gradient-descent technique is used to find the optimal weights to combine the individual kernels. Finally, SVM is used on the fused kernel matrix to classify the target

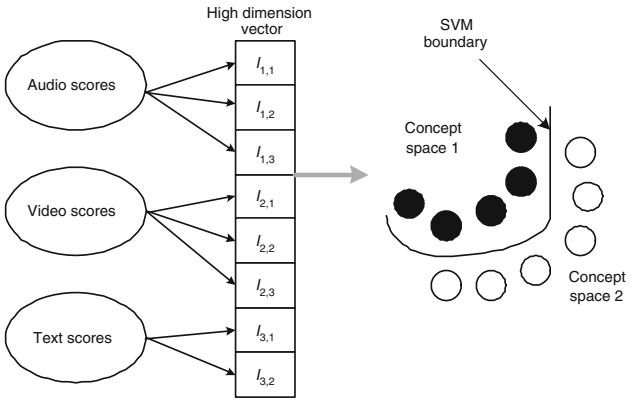

Fig. 3 SVM based score space classification of combined information from multiple intermediate concepts [3]

concept. Unlike GLF, the NLF method is used for non-linear combination of multimodal information. This method is based on [3], where SVM is first used as a classifier for the individual modality and then super kernel non-linear fusion is applied for optimal combination of the individual classifier models. The experiments on the TREC-2003 Video Track benchmark showed that NLF and GLF performed 8.0 and 5.0% better than the best single modality, respectively. Furthermore, NLF had an average 3.0% better performance than GLF. The NLF fusion approach was later extended by the authors [143] in order to obtain the best independent modalities (early fusion) and the strategy to fuse the best modalities (late fusion).

A hybrid fusion approach has been presented by Ayache et al. [11] as normalized early fusion and contextual late fusion for semantic indexing of multimedia resources using visual and text cues. Unlike other works, in case of normalized early fusion, each entry of the concatenated vector is normalized and then fused. In the case of contextual late fusion, the second layer classifier based on SVM is used to exploit the contextual relationship between the different concepts. Here, the authors have also presented a kernel-based fusion scheme based on SVMs, where the kernel functions are chosen according to the different modalities.

In the area of image classification, Zhu et al. [156] have reported a multimodal fusion framework to classify the images that have embedded text within their spatial coordinates. The fusion process followed two steps. At first, a bag-of-words model [73] is applied to classify the given image that considers the low-level visual features. In parallel, the text detector finds the text existence in the image using text color, size, location, edge density, brightness, contrast, etc. In the second step, a pair-wise SVM classifier is used for fusing the visual and textual features together. This is illustrated in Fig. 4.

In a recent work, Bredin and Chollet [19] proposed a biometric-based identification scheme of a talking face.

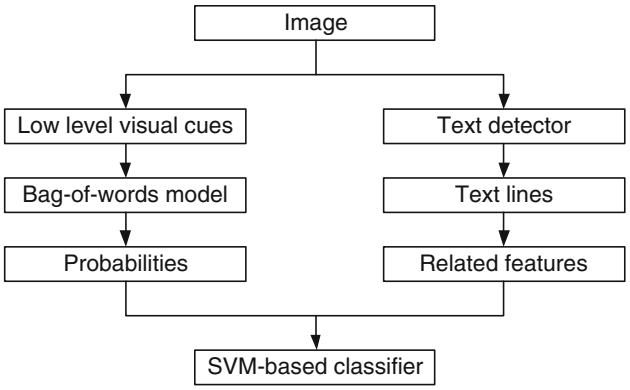

Fig. 4 Multimodal fusion using visual and text cues for image classification based on pair-wise SVM classifier [156]

The key idea was to utilize the synchrony measure between the talking face's voice and the corresponding video frames. Audio and visual sample rates are balanced by linear interpolation. By adopting a late fusion approach, the scores from the monomodal biometric speaker verification, face recognition, and synchrony were combined and passed to the SVM model, which provided the decision about the identity of the talking face. On another front, Aguilar et al. [4] provided a comparison between the rule-based fusion and learning-based fusion (trained) strategy. The scores of face, fingerprint and online signature are combined using both the Sum rule and radial basis function SVM (RBF SVM) for comparison. The experimental results demonstrates that learning-based RBF SVM scheme outperforms the rule-based scheme based on some appropriate parameter selection.

Snoek et al. [121] have compared both the early and late fusion strategies for semantic video analysis. Using the former approach, the visual vector has been concatenated with the text vector and then normalized to use as input in SVM to learn the semantic concept. In the latter approach the authors have adopted a probabilistic aggregation mechanism. Based on an experiment on 184 h of broadcast video using 20 semantic concepts, this study concluded that a late fusion strategy provided better performance for most concepts, but it bears an increased learning effort. The conclusion also suggested that when the early fusion performed better, the improvements were significant. However, which of the fusion strategies is better in which case needs further investigation.

3.2.2 *Bayesian inference*

The Bayesian inference is often referred to as the 'classical' sensor fusion method because it has been widely used and many other methods are based on it [45]. In this method, the multimodal information is combined as per the rules of probability theory [79]. The method can be applied at the feature level as well as at the decision level. The observations obtained from multiple modalities or the decisions obtained from different classifiers are combined, and an inference of the joint probability of an observation or a decision is derived [109].

The Bayesian inference fusion method is briefly described as follows. Let us fuse the feature vectors or the decisions $(I_1, I_2, \ldots, I_n)$ obtained from n different modalities. Assuming that these modalities are statistically independent, the joint probability of an hypothesis H based on the fused feature vectors or the fused decisions can be computed as [102]:

$$p(H|I_1, I_2, \ldots, I_n) = \frac{1}{N} \prod_{k=1}^{n} p(I_k|H)^{w_k} \quad (3)$$

where N is used to normalize the posterior probability estimate $p(H|I_1, I_2, \ldots, I_n)$. The term w_j is the weight of the kth modality, and $\sum_{k=1}^{n} w_j = 1$. This posterior probability is computed for all the possible hypotheses, E. The hypothesis that has the maximum probability is determined using the MAP rule $\hat{H} = \text{argmax}_{H \in E}\, p(H|I_1, I_2, \ldots, I_n)$.

The Bayesian inference method has various advantages. Based on the new observations, it can incrementally compute the probability of the hypothesis being true. It allows for any prior knowledge about the likelihood of the hypothesis to be utilized in the inference process. The new observation or the decision is used to update the a priori probability in order to compute the posterior probability of the hypothesis. Moreover, in the absence of empirical data, this method permits the use of a subjective probability estimate for the a priori of hypotheses [140].

These advantages of the Bayesian method are seen as its limitations in some cases. Bayesian inference method requires a priori and the conditional probabilities of the hypothesis to be well defined [110]. In absence of any knowledge of suitable priors, the method does not perform well. For example, in gesture recognition scenarios, it is sometimes difficult to classify a gesture of two stretched fingers in "V" form. This gesture can be interpreted as either "victory sign" or "sign indicating number two". In this case, since the priori probability of both the classes would be 0.5, the Bayesian method would provide ambiguous results. Another limitation of this method is that it is often found unsuitable for handling mutually exclusive hypotheses and general uncertainty. It means that only one hypothesis can be true at any given time. For example, Bayesian inference method would consider two events of human's running and walking mutually exclusive and cannot handle a fuzzy event of human's fast walking or slow running.

Bayesian inference method has been successfully used to fuse multimodal information (at the feature level and at the decision level) for performing various multimedia

analysis tasks. An example of Bayesian inference fusion at the feature level is the work by Pitsikalis et al. [102] for audio-visual speech recognition. Meyer et al. [85] and Xu and Chua [149] have used the Bayesian inference method at the decision level for spoken digit recognition and sports video analysis, respectively; while Atrey et al. [8] employed this fusion strategy at both the feature as well as the decision level for event detection in the multimedia surveillance domain. These works are described in the following.

Pitsikalis et al. [102] used the Bayesian inference method to combine the audio-visual feature vectors. The audio feature vector included 13 static MFCC and their derivatives, while the visual feature vector was formed by concatenating 6 shapes and 12 texture features. Based on the combined features, the joint probability of a speech segment is computed. In this work, the authors have also proposed to model the measurement of noise uncertainty.

At the decision level, Meyer et al. [85] fused the decisions obtained from speech and visual modalities. The authors have first extracted the MFCC features from speech and the lip contour features from the speaker's face in the video, and then obtained individual decisions (in terms of probabilities) for both using HMM classifiers. These probability estimates are then fused using the Bayesian inference method to estimate the joint probability of a spoken digit. Similar to this work, Xu and Chua [149] also used the Bayesian inference fusion method for integrating the probabilistic decisions about the offset and non-offset events detected in a sport video. These events have been detected by fusing audio-visual features with textual clues and by employing a HMM classifier. In this work, the authors have shown that the Bayesian inference has comparable accuracy to the rule-based schemes.

In another work, Atrey et al. [8] adopted a Bayesian inference fusion approach at hybrid levels (feature level as well as decision level). The authors demonstrated the utility of this fusion (they call it 'assimilation') approach for event detection in a multimedia surveillance scenario. The feature level assimilation was performed at the intra-media stream level and the decision level assimilation was adopted at the inter-media stream level.

3.2.3 *Dempster–Shafer theory*

Although the Bayesian inference fusion method allows for uncertainty modeling (usually by Gaussian distribution), some researchers have preferred to use the Dempster–Shafer (D–S) evidence theory since it uses belief and plausibility values to represent the evidence and their corresponding uncertainty [110]. Moreover, the D–S method generalizes the Bayesian theory to relax the Bayesian inference method's restriction on mutually exclusive hypotheses, so that it is able to assign evidence to the union of hypotheses [140].

The general methodology of fusing the multimodal information using the D–S theory is as follows. The D–S reasoning system is based on a fundamental concept of "the frame of discernment", which consists of a set Θ of all the possible mutually exclusive hypotheses. An hypothesis is characterized by *belief* and *plausibility*. The degree of belief implies a lower bound of the confidence with which a hypothesis is detected as true, whereas the plausibility represents the upper bound of the possibility that the hypothesis could be true. A probability is assigned to every hypothesis $H \in \mathcal{P}(\Theta)$ using a belief mass function $m : \mathcal{P}(\Theta) \rightarrow [0, 1]$. The decision regarding a hypothesis is measured by a "confidence interval" bounded by its basic belief and plausibility values, as shown in Fig. 5.

When there are multiple independent modalities, the D–S evidence combination rule is used to combine them. Precisely, the mass of a hypothesis H based on two modalities, I_i and I_j, is computed as:

$$(m_i \oplus m_j)(H) = \frac{\sum_{I_i \cap I_j = H} m_i(I_i) m_j(I_j)}{1 - \sum_{I_i \cap I_j = \varnothing} m_i(I_i) m_j(I_j)} \tag{4}$$

Note that, the weights can also be assigned to different modalities that are fused.

Although the Dempster–Shafer fusion method has been found more suitable for handling mutually inclusive hypotheses, this method suffers from the combinatorial explosion when the the number of frames of discernment is large [27].

Some of the representative works that have used the D–S fusion method for various multimedia analysis tasks are Bendjebbour et al. [16] (at hybrid level) and Mena and Malpica [84] (at the feature level) for segmentation of satellite images, Guironnet et al. [44] for video classification, Singh et al. [116] for finger print classification, and [110] for human computer interaction (at the decision level).

Bendjebbour et al. [16] proposed to use the D–S theory to fuse the mass functions of two regions (cloud and no

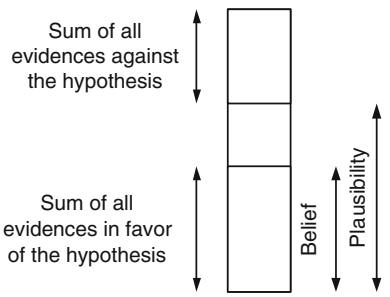

Fig. 5 An illustration of the belief and plausibility in the D–S theory [140]

cloud) of the image obtained from radar. They performed fusion at two levels the feature level and the decision level. At the feature level, the pixel intensity was used as a feature and the mass of a given pixel based on two sensors was computed and fused; while at the decision level, the decisions about a pixel obtained from the HMM classifier were used as mass and then the HMM outputs were combined. Similar to this work, Mena and Malpica [84] also used the D–S fusion approach for the segmentation of color images for extracting information from terrestrial, aerial or satellite images. However, they extracted the information of the same image from three different sources: the location of an isolated pixel, a group of pixels, and a pair of pixels. The evidences obtained based on the location analysis were fused using the D–S evidence fusion strategy.

Guironnet et al. [44] extracted low-level (color or texture) descriptors from a TREC video and applied a SVM classifier to recognize the pre-defined concepts (e.g. 'beach' or 'road') based on each descriptor. The SVM classifier outputs are integrated using the D–S fusion approach, they call it the "transferable belief model". In the area of biometrics, Singh et al. [116] used the D–S theory to combine the output scores of three different finger print classification algorithms based on the Minutiae, ridge and image pattern features. The authors showed that the D–S theory of fusing three independent evidences outperformed the individual approaches. Recently, Reddy [110] also used the D–S theory for fusing the outputs of two sensors, the Hand Gesture sensor and the Brain Computing Interface sensor. Two concepts, "Come" and "Here" were detected using these two sensors. The fusion results showed that the D–S fusion approach helps in resolving the ambiguity in the sensors.

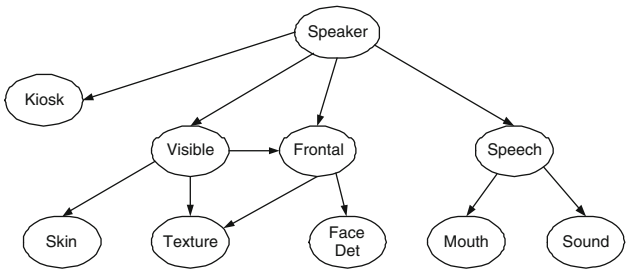

Fig. 6 An example of a static Bayesian networks [30]

3.2.4 *Dynamic Bayesian networks*

Bayesian inferencing can be extended to a network (graph) in which the nodes represent random variable (observations or states) of different types, e.g. audio and video; and the edges denote their probabilistic dependencies. For example, as shown in Fig. 6, a speaker detection problem can be depicted by a Bayesian Network [30]. The speaker node value is determined based on the value of three intermediate nodes 'visible', 'frontal' and 'speech', which are inferred from the measurement nodes 'skin', 'texture', 'face', 'mouth' and 'sound'. The figure shows the dependency of nodes upon each other. However, the network shown in Fig. 6 is a static one, meaning it depicts the state at a particular time instant.

A Bayesian Network works as a dynamic Bayesian network (DBN) when the temporal aspect being added to it as shown in Fig. 7. A DBN is also called a *probabilistic generative model* or a *graphical model*. Due to the fact that

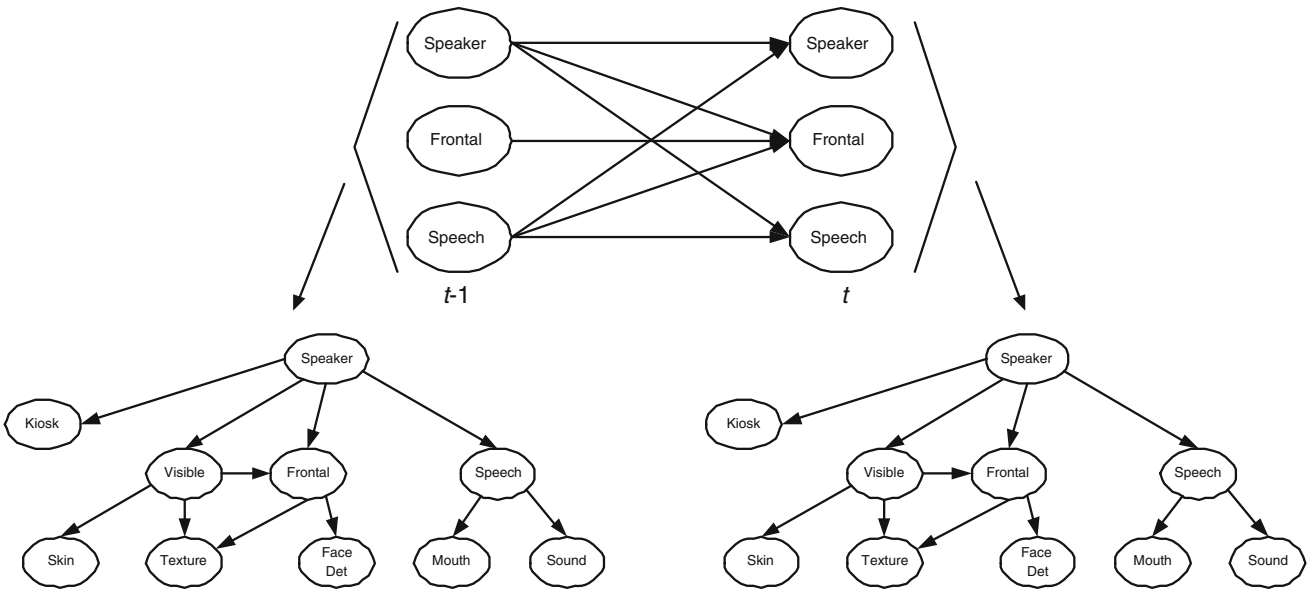

Fig. 7 An example of a dynamic Bayesian networks [30]

these models describe the observed data in terms of the process that generate them, they are called generative. They are termed as probabilistic because they describe probabilistic distributions rather than the sensor data. Moreover, since they have useful graphical representations, they are also called graphical [15]. Although the DBNs have been used with different names such as probabilistic generative models, graphical models, etc. for a variety of applications, the most popular and simplest form of a DBN is the HMM.

The DBNs have a clear advantage over the other methods in two aspects. First, they are capable of modeling the multiple dependencies among the nodes. Second, by using them, the temporal dynamics of multimodal data can easily be integrated [119]. These advantages make them suitable for various multimedia analysis tasks that require decisions to be performed using time-series data. Although DBNs are very beneficial and widely used, the determination of the right DBN state is often seen as its problem [81].

In the following, we briefly outline some representative works that have used DBN in one form or the other. Wang et al. [138] have used HMM for video shot classification. The authors have extracted both audio (cepstral vector) and visual features (a gray-level histogram difference and two motion features) from each video frame and used them as the input data for the HMM. While this method used a single HMM that processed the joint audio-visual features, Nefian et al. [86] used the coupled HMM (CHMM), which is a generalization of the HMM. The CHMM suits to multimodal scenarios where two or more streams need to be integrated. In this work, the authors have modeled the state asynchrony of the audio features (MFCC) and visual features (2D-DCT coefficients of the lips region) while preserving their correlation over time. This approach is used for speech recognition. The work by Adams et al. [3] also used a Bayesian network in addition to SVM and showed the comparison of both for video shot retrieval.

Unlike Nefian et al. [86] who used CHMM, Bengio [17] has presented the asynchronous HMM (AHMM) at the feature level. The AHMM is a variant of HMM to deal with the asynchronous data streams. The authors modeled the joint probability distribution of asynchronous sequences—speech (MFCC features) stream and video (shape and intensity features) stream that described the same event. This method was used for biometric identity verification.

Nock et al. [90] and [91] employed a set of HMMs trained on joint sequences of audio and visual data. The features used were MFCC from speech and DCT coefficients of the lip region in the face from video. The joint features were presented to the HMMs at consecutive time instances in order to locate a speaker. The authors also computed the mutual information (MI) between the two types of features and analyzed its effect on the overall speaker location results. Similar to this work, Beal et al. [15] have used graphical models to fuse audio-visual observations for tracking a moving object in a cluttered, noisy environment. The authors have modeled audio and video observations jointly by computing their mutual dependencies. The expectation-maximization algorithm has been used to learn the model parameters from a sequence of audio-visual data. The results were demonstrated in a two microphones and one camera setting. Similarly, Hershey et al. [46] also used a probabilistic generative model to combine audio and video by learning the dependencies between the noisy speech signal from a single microphone and the fine-scale appearance and location of the lips during speech.

It is important to note that all works described above assume multiple modalities (usually audio-visual data) that locally, as well as jointly, follow a Gaussian distribution. In contrast to these works, Fisher et al. [39] have presented a non-parametric approach to learn the joint distribution of audio and visual features. They estimated a linear projection onto low-dimensional subspaces to maximize the mutual information between the mapped random variables. This approach was used for audio–video localization. Although the non-parametric approach is free from any parametric assumptions, they often suffer from implementation difficulties as the method results in a system of undetermined equations. That is why, the parametric approaches have been preferred [92]. With this rationale, Noulas and Krose [92] also presented a two-layer Bayesian network model for human face tracking. In the first layer, the independent modalities (audio and video) are analyzed, while the second layer performs the fusion incorporating their correlation. A similar approach was also presented by Zou and Bhanu [158] for tracking humans in a cluttered environment. Recently, Town [131] also used the Bayesian networks approach for multi-sensory fusion. In this work, the visual information obtained from the calibrated cameras is integrated with the ultrasonic sensor data at the decision level to track people and devices in an office building. The authors presented a large-scale sentient computing system known as "SPIRIT".

In the context of news video analysis, Chua et al. [31] have emphasized on the need to utilize multimodal features (text with audio-visual) for segmenting news video into story units. In their other work [25], the authors presented an HMM-based multi-modal approach for news video story segmentation by using a combination of features. The feature set included visual-based features such as color, object-based features such as face, video-text, temporal features such as audio and motion, and semantic features such as cue-phrases. Note that the fundamental assumption

which is often considered with the DBN methods is the independence among different observations/features. However, this assumption does not hold true in reality. To relax the assumption of independence between observations, Ding and Fan [38] presented a segmental HMM approach to analyze a sports video. In segmental HMM, each hidden state emits a sequence of observations, which is called a segment. The observations within a segment are considered to be independent to the observations of other segments. The authors showed that the segmental HMM performed better than traditional HMM. In another work, the importance of combining text modality with the other modalities has been demonstrated by Xie et al. [145]. The authors proposed a layered dynamic mixture model for topic clustering in video. In their layer approach, first a hierarchical HMM is used to find clusters in audio and visual streams; and then latent semantic analysis is used to cluster the text from the speech transcript stream. At the next level, a mixture model is adopted to learn the joint probability of the clusters from the HMM and latent semantic analysis. The authors have performed experiments with the TRECVID 2003 data set, which demonstrated that the multi-modal fusion resulted in a higher accuracy in topics clustering.

An interesting work was presented by Wu et al. [142]. In this work, the authors used an influence diagram approach (a form of the Bayesian network) to represent the semantics of photos. The multimodal fusion framework integrated the context information (location, time and camera parameters), content information (holistic and perceptual local features) with the domain-oriented semantic ontology (represented by a directed acyclic graph). Moreover, since the conditional probabilities that are used to infer the semantics can be misleading, the authors have utilized the causal strength between the context/content and semantic ontology instead of using the correlation among features. The causal strength is based on the following idea. The two variables may co-vary with each other, however, there may be a third variable as a "cause" that may affect the value of these two variables. For example, two variables "wearing a warm jacket" and "drinking coffee" may have a large positive correlation; however, the cause behind both could be "cold weather". The authors have shown that the usage of causal strength in influence diagrams provide better results in the automatic annotation of photos.

3.2.5 *Neural networks*

Neural network (NN) is another approach for fusing multimodal data. Neural networks are considered a non-linear black box that can be trained to solve ill-defined and computationally expensive problems [140]. The NN method consists of a network of mainly three types of nodes—input, hidden and output nodes. The input nodes accept sensor observations or decisions (based on these observations), and the output nodes provide the results of fusion of the observations or decisions. The nodes that are neither input nor output are referred to hidden nodes. The network architecture design between the input and output nodes is an important factor for the success or failure of this method. The weights along the paths, that connect the input nodes to the output nodes, decide the input–output mapping behavior. These weights can be adjusted during the training phase to obtain the optimal fusion results [22]. This method can also be employed at both the feature level and the decision level.

In the following, we describe some works that illustrate the use of the NN fusion method for performing the multimedia analysis tasks. Gandetto et al. [41] have used the NN fusion method to combine sensory data for detecting human activities in an environment equipped with a heterogeneous network of sensors with CCD cameras and computational units working together in a LAN. In this work, the authors considered two types of sensors—state sensors (e.g. CPU load, login process, and network load) and observation sensors (e.g. cameras). The human activities in regard to usage of laboratory resources were detected by fusing the data from these two types sensors at the decision level.

A variation of the NN fusion method is the time-delay neural network (TDNN) that has been used to handle temporal multimodal data fusion. Some researchers have adopted the TDNN approach for various multimedia analysis tasks, e.g. Cutler and Davis [34], Ni et al. [88], and Zou and Bhanu [158] for speaker tracking. Cutler and Davis [34] learned the correlation between audio and visual streams by using a TDNN method. The authors have used it for locating the speaking person in the scene. A similar approach was also presented by Zou and Bhanu [158]. In [158], the authors have also compared the TDNN approach with the BN approach and found that the BN approach performed better than the TDNN approach in many aspects. First, the choosing of the initial parameters does not affect the DBN approach while it does affect the TDNN approach. Second, the DBN approach was better in modeling the joint Gaussian distribution of audio-visual data compared to linear mapping between audio signals and the object position in video sequences in the TDNN method. Third, the graphical models provide an explicit and easily accessible structure compared to TDNN, in which, the inner structure and parameters are difficult to design. Finally, the DBN approach offers better tracking accuracy. Moreover, in the DBN approach, a posteriori probability of the estimates is available as the quantitative measure in the support of the decision.

While Cutler and Davis [34] and Zou and Bhanu [158] used the NN fusion approach at the feature level, Ni et al. [88] adopted this approach at the feature level as well as at the decision level. In [88], the authors have used the NN fusion method to fuse low level features to recognize images. The decisions from multiple trained NN classifiers are further fused to come up with a final decision about an image.

Although the NN method is in general found suitable to work in a high-dimensional problem space and generate high-order nonlinear mapping, there are some familiar complexities associated with them. For instance, the selection of appropriate network architecture for a particular application is often difficult. Moreover, this method also suffers from slow training. Due to these limitations and other shortcomings (stated above as compared to the BN method), the NN method has not been often used for multimedia analysis tasks compared to other fusion methods.

3.2.6 Maximum entropy model

In general, maximum entropy model is a statistical classifier which follows an information-theoretic approach and provides a probability of an observation belonging to a particular class based on the information content it has. This method has been used by few researchers for categorizing the fused multimedia observations into respective classes.

The maximum entropy model based fusion method is briefly described as follows. Let I_i and I_j are the two different types of input observations. The probability of these observations belonging to a class X can be given by an exponential function:

$$P(X|I_i, I_j) = \frac{1}{Z(I_i, I_j)} e^{F(I_i, I_j)} \quad (5)$$

where, $F(I_i, I_j)$ is the combined feature (or decision) vector and $Z(I_i, I_j)$ is the normalization factor to ensure a proper probability.

Recently, this fusion method has been used by Magalhães and Rüger [80] for semantic multimedia indexing. In this work, the authors combined the text and image based features to retrieve the images. The authors found that the maximum entropy model based fusion worked better than the Naive Bayes approach.

There are other works such as Jeon and Manmatha [63] and Argillander et al. [7], which have used maximum entropy model for multimedia analysis tasks, however in these works the authors used only single modality rather than multiple modalities. Therefore, the discussion of these works is out of scope for our paper.

3.2.7 Remarks on classification-based fusion methods

All the representative works related to the classification-based fusion methods are summarized in Table 2. Our observations are as follows:

- The Bayesian inference fusion method, which works on probabilistic principles, provides easy integration of new observation and the use of a priori information. However, they are not suitable for handling mutually exclusive hypotheses. Moreover, the lack of appropriate a priori information can lead to inaccurate fusion results using this method. On the other hand, Dempster–Shafer fusion methods are good at handling certainty and mutually exclusive hypotheses. However, in this method, it is hard to handle the large number of combinations of hypotheses. This method has been used for speech recognition, sports video analysis and event detection tasks.
- The dynamic Bayesian networks have been widely used to deal with time-series data. This method is a variation of the Bayesian Inference when used over time. The DBN method in its different forms (such as HMMs) has been successfully used for various multimedia analysis tasks such as speech recognition, speaker identification and tracking, video shot classification etc. However, in this method, it is often difficult to determine the right DBN states. Compared to DBN, the neural networks fusion method is generally suitable to work in a high-dimensional problem space and it generates a high-order nonlinear mapping, which is required in many realistic scenarios. However, due to the complex nature of a network, this method suffers from slow training.
- As can be seen from the table, among various classification-based fusion methods, SVM and DBN have been widely used by researchers. SVMs have been preferred due to their improved classification performance while the DBNs have been found more suitable to model temporal data.
- There are various other classification methods used in multimedia research. These include decision tree [76], relevance vector machines [36], logistics regression [71] and boosting [75]. However, these methods have been used more for the traditional classification problems than for the fusion problems. Hence, we skip the description of these methods.

3.3 Estimation-based fusion methods

The estimation category includes the Kalman filter, extended Kalman filter and particle filter fusion methods. These methods have been primarily used to better estimate

Table 2 A list of the representative works in the classification methods category used for multimodal fusion

Fusion method	Level of fusion	The work	Modalities	Multimedia analysis task
Support vector machine	Decision	Adams et al. [3]	Video (color, structure, and shape), audio (MFCC) and textual cues	Semantic concept detection
		Aguilar et al. [4]	Fingerprint, signature, face (MCYT Multimodal Database, XM2VTS face database)	Biometric verification
		Iyenger et al. [58]	Audio, video	Semantic concept detection
		Wu et al. [141]	Color histogram, edge orientation histogram, color correlogram, co-occurrence texture, motion vector histogram, visual perception texture, and speech	Semantic concept detection
		Bredin and Chollet [19]	Audio (MFCC), video (DCT of lip area), audio-visual speech synchrony	Biometric identification of talking face
	Hybrid	Wu et al. [143]	Video, audio	Multimedia data analysis
		Zhu et al. [156]	Image (low-level visual features, text color, size, location, edge density, brightness, contrast)	Image classification
		Ayache et al. [11]	Visual, text cue	Semantic indexing
Bayesian inference	Feature	Pitsikalis et al. [102]	Audio (MFCC), video (Shape and texture)	Speech recognition
	Decision	Meyer et al. [85]	Audio (MFCC) and video (lips contour)	Spoken digit recognition
	Hybrid	Xu and Chua [149]	Audio, video, text, web log	Sports video analysis
		Atrey et al. [8]	Audio (ZCR, LPC, LFCC) and video (blob location and area)	Event detection for surveillance
Dempster–Shafer theory	Feature	Mena and Malpica [84]	Video (trajectory coordinates)	Segmentation of satellite images
	Decision	Guironnet et al. [44]	Audio (phonemes) and visual (visemes)	Video classification
		Singh et al. [116]	Audio (MFCC), video (DCT of the face region) and the synchrony score	Finger print classification
		Reddy [110]	Audio (MFCC), video (Eigenlip)	Human computer interaction
	Hybrid	Bendjebbour et al. [16]	Video (trajectory coordinates)	Segmentation of satellite images
Dynamic Bayesian networks	Feature	Wang et al. [138]	Audio (cepstral vector), visual (gray-level histogram difference and motion features)	Video shot classification
		Nefian et al. [86]	Audio (MFCC) and visual (2D-DCT coefficients of the lips region)	Speech recognition
		Nock et al. [90, 91]	Audio (MFCC) and video (DCT coefficients of the lips region	Speaker localization
		Chaisorn et al. [25]	Audio (MFCCs and perceptual features), video (color, face, video-text, motion)	Story segmentation in news video
		Adams et al. [3]	Video (color, structure, and shape),audio (MFCC) and textual cues	Video shot classification
		Beal et al. [15]	Audio and video—the details of features not available	Object tracking
		Bengio et al. [17]	Speech (MFCC) and video (shape and intensity features)	Biometric identity verification
		Hershey et al. [46]	Audio (Spectral components), video (fine-scale appearance and location of the lips)	Speaker localization
		Zou and Bhanu [158], Noulas and Krose [92]	Audio (MFCC) and video (pixel value variation)	Human tracking
		Ding and Fan [38]	Video (spatial color distribution and the angle of yard lines)	Shot classification in a sports video

Table 2 continued

Fusion method	Level of fusion	The work	Modalities	Multimedia analysis task
	Decision	Wu et al. [142]	Image (color, texture and shape) and Camera parameters	Photo annotation
		Town [131]	Video (face and blob), ultrasonic sensors	Human tracking
	Hybrid	Xie et al. [145]	Text (closed caption), Audio (pitch, silence, significant pause), video (color histogram and motion intensity), Speech (ASR transcript)	Topic clustering in video
Neural networks	Feature	Cutler and Davis [34]	Audio (phoneme) and video (viseme)	Speaker localization
		Zou and Bhanu [158]	Audio (spectrogram) and video (blob)	Human tracking
	Decision	Gandetto et al. [41]	CPU load, Login process, Network load, Camera images	Human activity monitoring
	Hybrid	Ni et al. [88]	Image (features details not provided in the paper)	Image recognition
Maximum Entropy Model	Feature	Magalhães and Rüger [80]	Text and Image	Semantic image indexing

the state of a moving object based on multimodal data. For example, for the task of object tracking, multiple modalities such as audio and video are fused to estimate the position of the object. The details of these methods are as follows.

3.3.1 *Kalman filter*

The Kalman filter (KF) [66, 112] allows for real-time processing of dynamic low-level data and provides state estimates of the system from the fused data with some statistical significance [79]. For this filter to work, a linear dynamic system model with Gaussian noise is assumed, where at time t, the system true state, $x(t)$ and its observation, $y(t)$ are modeled based on the state at time $t-1$. Precisely, this is represented using the state-space model given by Eqs. 6 and 7 in the following:

$$x(t) = A(t)x(t-1) + B(t)I(t) + w(t) \tag{6}$$

$$y(t) = H(t)x(t) + v(t) \tag{7}$$

where, $A(t)$ is the transition model, $B(t)$ is the control input model, $I(t)$ is the input vector, $H(t)$ is the observation model, $w(t) \sim N(0, Q(t))$ is the process noise as a normal distribution with zero mean and $Q(t)$ covariance, and $v(t) \sim N(0, R(t))$ is the observation noise as a normal distribution with zero mean and $R(t)$ covariance.

Based on the above state-space model, the KF does not require to preserve the history of observation and only depends on the state estimation data from the previous timestamp. The benefit is obvious for systems with less storage capabilities. However, the use of the KF is limited to the linear system model and is not suitable for the systems with non-linear characteristics. For non-linear system models, a variant of the Kalman filter known as extended Kalman filter (EKF) [111] is usually used. Some researchers also use KF as inverse Kalman filter (IKF) that reads an estimate and produce an observation as oppose to KF that reads an observation and produce an estimate [124]. Therefore, a KF and its associated IKF can logically be arranged in series to generate the observation at the output. Another variant of KF has gained attention lately, which is the unscented Kalman filter (UKF) [65]. The benefit of UKF is that it does not have a linearization step and the associated errors.

The KF is a popular fusion method. Loh et al. [77] proposed a feature level fusion method for estimating the translational motion of a single speaker. They used different audio-visual features for estimating the position, velocity and acceleration of the single sound source. For position estimation in 3D space, the measurement of three microphones are used in conjunction with the camera image point. Given the position estimate, a KF is then

used based on a constant acceleration model to estimate velocity and acceleration. Unlike Loh et al. [77], Potamitis et al. [107] presented a audio-based fusion scheme for detecting multiple moving speakers. The same speaker state is determined by fusing the location estimates from multiple microphone arrays, where the location estimates are computed using separate KF for all the individual microphone arrays. A probabilistic data association technique is used with an interacting multiple model estimator to handle speaker's motion and measurement origin uncertainty.

KF as well as EKF have been successfully used for source localization and tracking for many years. Strobel et al. [124] focused on the localization and tracking of single objects. The audio and video localization features are computed in terms of position estimates. EKF is used due to the non-linear estimates based on audio-based position. On the other hand, a basic KF is used at the video camera level. The outputs of the audio and video estimates are then fused within the fusion center, which is comprised of two single-input inverse KFs and a two-input basic KFs. This is shown in Fig. 8. This work requires that the audio and video sources are in sync with each other. Likewise, Talantzis et al. [125] have adopted a decentralized KF that fuses audio and video modalities for better location estimation in real time. A decision level fusion approach has been adopted in this work.

A recent work [154] presents a multi-camera based tracking system, where multiple features such as spatial position, shape and color information are integrated together to track object blobs in consecutive image frames. The trajectories from multiple cameras are fused at feature level to obtain the position and velocity of the object in the real world. The fusion of trajectories from multiple cameras, which uses EKF, enables better tracking even when the object view is occluded. Gehrig et al. [43] also adopted an EKF based fusion approach using audio and video features. Based on the observation of the individual audio and video sensor, the state of the KF was incrementally updated to estimate the speaker's position.

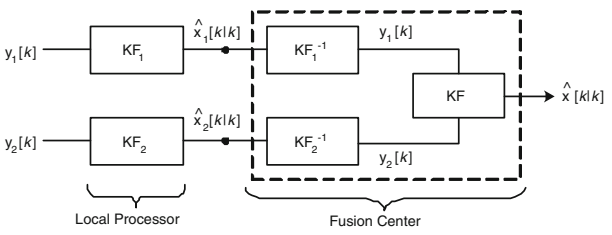

Fig. 8 Extended/decentralized Kalman filter in the fusion process [124]

3.3.2 *Particle filter*

Particle filters are a set of sophisticated simulation-based methods, which are often used to estimate the state distribution of the non-linear and non-Gaussian state-space model [6]. These methods are also known as Sequential Monte Carlo (SMC) methods [33]. In this approach, the particles represent the random samples of the state variable, where each particle is characterized by an associated weight. The particle filtering algorithm also consists of a prediction and update steps. The prediction step propagates each particle as per its dynamics while the update step reweighs a particle according to the latest sensory information. While the KF, EKF or IKF are optimal only for linear Gaussian processes, the particle methods can provide Bayesian optimal estimates for non-linear non-Gaussian processes when sufficiently large number of samples are taken.

The particle methods have been widely used in multimedia analysis. For instance, Vermaak et al. [132] used particle filters to estimate the predictions from audio- and video-based observations. The reported system uses a single camera and a pair of microphones and were tested based on stored audio-visual sequences. The fusion of audio-visual features took place at the feature level, meaning that the individual particle coordinates from the features of both modalities were combined to track the speaker. Similar to this approach, Perez et al. [99] adopted the particle filter approach to fuse 2D object shapes and audio information for speaker tracking. However, unlike Vermaak et al. [132], the latter uses the concept of importance particle filter, where audio information was specifically used to generate an importance function that influenced the computation of audio-based observation likelihood. The audio and video-based observation likelihoods are then combined as a late fusion scheme using a standard probabilistic product formula that forms the multimodal particle.

A probabilistic particle filter framework is proposed by Zotkin et al. [157] that adopts a late fusion approach for tracking people in a videoconferencing environment. This framework used multiple cameras and microphones to estimate the 3D coordinates of the person using the sampled projection. Multimodal particle filters are used to approximate the posterior distribution of the system parameters and the tracking position in the audio-visual state-space model. Unlike Vermaak et al. [132] or Perez et al. [99], this framework enables tracking multiple persons simultaneously.

Nickel et al. [89] presented an approach for real-time tracking of the speaker using multiple cameras and microphones. This work used particle filters to estimate the location of the speaker by sampled projection as proposed

by Zotkin et al. [157], where each particle filter represented a 3D coordinate in the space. The evidence from all the camera views and microphones are adjusted to assign weights to the corresponding particle filter. Finally, the weighted mean of a particle set is considered as the speaker location. This work adopted a late fusion approach to obtain the final decision.

3.3.3 Remarks on estimation-based fusion methods

The representative works in the estimation-based category are summarized in Table 3. The estimation-based fusion methods (Kalman filter, extended Kalman filter and particle filter) are generally used to estimate and predict the fused observations over a period. These methods are suitable for object localization and tracking tasks. While the Kalman filter is good for the systems with a linear model, the extended Kalman filter is better suited for non-linear systems. However, the particle filter method is more robust for non-linear and non-Gaussian models as they approach the Bayesian optimal estimate with sufficiently large number of samples.

3.4 Further discussion

In the following, we provide our observations based on the analysis of the fusion methods described above.

- *Most used methods*. From the literature, it has been observed that many fusion methods such as linear weighted fusion, SVM, and DBN have been used more often in comparison to the other methods. This is due to the fact the linear weighted fusion can be easily used to prioritize different modalities while fusing; SVM has improved classification performance in many multimedia analysis scenarios; and the DBN fusion method is capable of handling temporal dependencies among multimodal data, which is an important issue often considered in multimodal fusion.
- *Fusion methods and levels of fusion*. Existing literature suggest that linear weighted fusion is suitable to work at the decision level. Also, although SVM is generally used to classify individual modalities at the feature level, in the case of multimodal fusion, the outputs of individual SVM classifiers are fused and further classified using another SVM. That is why most of the reported works have been seen to fall into the late fusion category. Among others, the DBNs have been used more at the feature level due to its suitability in handling temporal dependencies.
- *Modalities used*. The modalities that have been used for multimodal fusion are mostly based on audio and video. Some works also considered text modality, while others have investigated gesture.
- *Multimedia analysis tasks versus fusion methods*. In Table 4, we summarize the existing literature in terms of the multimedia analysis tasks and the different fusion methods used for these tasks. This may be useful for the readers as a quick reference in order to decide which fusion method would be suitable for which task. It has been found that for a variety of tasks, various fusion methodologies have been adopted. However, based on the nature of a multimedia analysis task, some fusion methods have been preferred over the others. For

Table 3 A list of the representative works in the estimation methods category used for multimodal fusion

Fusion method	Level of fusion	The work	Modalities	Multimedia analysis task
Kalman filter and its variants	Feature	Potamitis et al. [107]	Audio (position, velocity)	Multiple speaker tracking
		Loh et al. [77]	Audio, video	Single speaker tracking
		Gehrig et al. [43]	Audio (TDOA), video (position of the speaker)	Single speaker tracking
		Zhou and Aggarwal [154]	Video [spatial position, shape, color (PCA), blob]	Person/vehicle tracking
	Decision	Strobel et al. [124]	Audio, video	Single object localization and tracking
		Talantzis et al. [125]	Audio (DOA), video (position, velocity, target size)	Person tracking
Particle filter	Feature	Vermaak et al. [132]	Audio (TDOA), visual (gradient)	Single speaker tracking
	Decision	Zotkin et al. [157]	Audio (TDOA), video (skin color, shape matching and color histograms)	Multiple speaker tracking
		Perez et al. [99]	Audio (TDOA), video (coordinates)	Single speaker tracking
		Nickel et al. [89]	Audio (TDOA), video (Haar-like features)	Single speaker tracking

Table 4 A summary of the fusion method used for different multimedia analysis tasks

Multimedia analysis task	Fusion method	The works
Biometric identification and verification	Support vector machine	Bredin and Chollet [19], Aguilar et al. [4]
	Dynamic Bayesian networks	Bengio et al. [17]
Face detection, human tracking and activity/event detection	Linear weighted fusion	Kankanhalli et al. [67], Jaffre and Pinquier [59]
	Bayesian inference	Atrey et al. [8]
	Dynamic Bayesian networks	Town [131], Beal et al. [15]
	Neural networks	Gandetto et al. [41], Zou and Bhanu [158]
	Kalman filter	Talantzis et al. [125], Zhou and Aggarwal [154], Strobel et al. [124]
Human computer interaction and multimodal dialog system	Custom-defined rules	Corradini et al. [32], Pfleger [100], Holzapfel et al. [49]
	Dempster–Shafer theory	Reddy [110]
Image segmentation, classification, recognition, and retrieval	Linear weighted fusion	Hua and Zhang [55]
	Support vector machine	Zhu et al. [156]
	Neural networks	Ni et al. [88]
	Dempster–Shafer Theory	Mena and Malpica [84], Bendjebbour et al. [16]
Video classification and retrieval	Linear weighted fusion	Yan et al. [151], McDonald and Smeaton [83]
	Bayesian inference	Xu and Chua [149]
	Dempster–Shafer Theory	Singh et al. [116]
	Dynamic Bayesian networks	Wang et al. [138], Ding and Fan [38], Chaisorn et al. [25], Xie et al. [145], Adams et al. [3]
Photo and video annotation	Linear weighted fusion	Iyenger et al. [58]
	Dynamic Bayesian networks	Wu et al. [142]
Semantic concept detection	Linear weighted fusion	Iyenger et al. [58]
	Support vector machine	Adams et al. [3], Iyenger et al. [58], Wu et al. [141]
Semantic multimedia indexing	Custom-defined rules	Babaguchi et al. [12]
	Support vector machine	Ayache et al. [11]
	Maximum entropy model	Magalhães and Rüger [80]
Monologue detection	Linear weighted fusion	Kankanhalli et al. [67], Iyenger et al. [57]
Speaker localization and tracking	Particle filter	Vermaak et al. [132], Perez et al. [99], Nickel et al. [89], Zotkin et al. [157]
	Kalman filter	Potamitis et al. [107]
	Majority voting rule	Radova and Psutka [108]
	Dynamic Bayesian networks	Nock et al. [91], Hershey et al. [46]
	Neural networks	Cutler and Davis [34]
Speech and speaker recognition	Linear weighted fusion	Neti et al. [87], Lucey et al. [78]
	Bayesian inference	Meyer et al. [85], Pitsikalis et al. [102]
	Dynamic Bayesian networks	Nefian et al. [86]

instance, for image and video classification retrieval tasks, the classification-based fusion methods such as Bayesian inference, Dempster–Shafer theory and dynamic Bayesian networks have been used. Also, as an object tracking task involves dynamics and the state transition and estimation, dynamic Bayesian networks and the estimation-based methods such as Kalman filter have widely been found successful. Moreover, since the sports and news analysis tasks consist of complex rules, custom-defined rules have been found appropriate.

- *Application constraints and the fusion methods*. From the perspective of application constraints such computation, delay and resources, we can analyze different fusion methods as follows. It has been observed that the

linear weighted fusion method is applied to the applications which have lesser computational needs. On other hand, while the dynamic Bayesian networks fusion method is computationally more expensive than the others, the neural networks can be trained to computationally expensive problems. Regarding the time delay and synchronization problems, custom-defined rules have been found more appropriate as they are usually application specific. These time delays may occur due to the resource constraints since the input data can be obtained from different types of multimedia sensors and the different CPU resources may be available for analysis.

4 Distinctive issues of multimodal fusion

This section provides a critical look at the distinctive issues that should be considered in a multimodal fusion process. These issues have been identified in the light of the following three aspects of fusion: how to fuse (in continuation with the fusion methodologies as discussed in Sect. 3), when to fuse, and what to fuse. From the aspect of *how to fuse*, we will elaborate in Sect. 4.1 on the issues of the use of correlation, confidence and the contextual information while fusing different modalities. The *when to fuse* aspect is related to the synchronization between different modalities which will be discussed in Sect. 4.2. We will cover *what to fuse* aspect by describing the issue of optimal modality selection in Sect. 4.3. In the following, we highlight the importance of considering these distinctive issues and also describe the past works related to them.

4.1 Issues related to how to fuse

4.1.1 Correlation between different modalities

The correlation among different modalities represents how they co-vary with each other. In many situations, the correlation between them provides additional cues that are very useful in fusing them. Therefore, it is important to know different methods of computing correlations and to analyze them from the perspective of how they affect fusion [103].

The correlation can be comprehended at various levels, e.g. the correlation between low level features and the correlation between semantic-level decisions. Also, there are different forms of correlation that have been utilized by the researchers in the multimodal fusion process. The correlation between features has been computed in the forms of correlation coefficient, mutual information, latent semantic analysis (also called lament semantic indexing), canonical correlation analysis, and cross-modal factor analysis. On the other hand, the decision level correlation has been exploited in the form of causal link analysis, causal strength and agreement coefficient.

Table 5 A list of some representative works that used the correlation information between different streams in the fusion process

Level of fusion	The form of correlation	The works	Multimedia analysis task
Feature	Correlation coefficient	Wang et al. [138]	Video shot classification
		Nefian et al. [86]	Speech recognition
		Beal et al. [15]	Object tracking
		Li et al. [74]	Talking face detection
	Mutual information	Fisher-III et al. [39]	Speech recognition
		Darrell et al. [35]	Speech recognition
		Hershey and Movellan [47]	Speaker localization
		Nock et al. [90], Iyengar et al. [57]	Monologue detection
		Nock et al. [91]	Speaker localization
		Noulas and Krose [92]	Human tracking
	Latent semantic analysis	Li et al. [74]	Talking face detection
		Chetty and Wagner [28]	Biometric person authentication
	Canonical correlation analysis	Slaney and Covell [117]	Talking face detection
		Chetty and Wagner [28]	Biometric person authentication
		Bredin and Chollet [20]	Talking face identity verification
	Cross-modal factor analysis	Li et al. [72]	Talking head analysis
Decision	Casual link analysis	Stauffer [123]	Event detection for surveillance
	Causal strength	Wu et al. [142]	Photo annotation
	Agreement coefficient	Atrey et al. [8]	Event detection for surveillance

In the following, we describe the above eight forms of correlation and their usage for the various multimedia analysis tasks. We also cast light on the cases where independence between different modalities can be useful for multimedia analysis tasks. A summary of the representative works that have used correlation in different forms is provided in Table 5.

Correlation coefficient. The correlation coefficient is a measure of the strength and direction of a linear relationship between any two modalities. It has been widely used by multimedia researchers for joint modeling the audio–video relationship [138, 86, 15]. However, to jointly model the audio–video, the authors have often assumed them—(1) to be independent, and (2) to locally and jointly follow the Gaussian distribution.

One of the most simple and widely used forms of the correlation coefficient is the Pearson's product-moment coefficient [20], which is computed as follows. Assuming that I_i and I_j are the two modalities (of same or different types). The correlation coefficient $\mathrm{CC}(I_i, I_j)$ between them can be computed as [138]:

$$\mathrm{CC}(I_i, I_j) = \frac{\hat{C}(I_i, I_j)}{\sqrt{\hat{C}(I_i, I_i)}\sqrt{\hat{C}(I_j, I_j)}} \tag{8}$$

where $\hat{C}(I_i, I_j)$ is the (i, j)th element of the covariance matrix C, which is given as:

$$C = \sum_{k=1}^{N} (I_i^k - I_i^m) \times (I_j^k - I_j^m) \tag{9}$$

where I_i^k and I_j^k are the kth value in the feature vector I_i and I_j, respectively; and I_i^m and I_j^m are the mean values of these feature vectors. This method of computing correlation has been used by many researchers such as Wang et al. [138] and Li et al. [74]. In [74], based on the Correlation Coefficient between audio and face feature vectors, the authors have selected the faces having the maximum correlation with the audio.

Mutual information. The mutual information is a information theoretic measure of correlation that represents the amount of information one modality conveys about another. The mutual information $\mathrm{MI}(I_i, I_j)$ between two modalities I_i and I_j, which are normally distributed with variances Σ_{I_i} and Σ_{I_j}, and jointly distributed with covariance $\Sigma_{I_i I_j}$, is computed as:

$$\mathrm{MI}(I_i, I_j) = \frac{1}{2}\log\frac{|\Sigma_{I_i}||\Sigma_{I_j}|}{|\Sigma_{I_i I_j}|} \tag{10}$$

There are several works that have used mutual information as a measure of synchrony between audio and video. For instance, Iyengar et al. [57] computed the synchrony between face and speech using mutual information. The authors found that the face region had high mutual information with the speech data. Therefore, the mutual information score helped locate the speaker. Similarly, Fisher et al. [39] also learned the linear projections from a joint audio–video subspace where the mutual information was maximized. Other works that have used mutual information as a measure of correlation are Darrell et al. [35], Hershey and Movellan [47], Nock et al. [90], Nock et al. [91] and Noulas and Krose [92] for different tasks as detailed in Table 5.

Latent semantic analysis. Latent semantic analysis (LSA) is a technique often used for text information retrieval. This technique has proven useful to analyze the semantic relationships between different textual units. In the context of text information retrieval, the three primary goals that the LSA technique achieves are dimension reduction, noise removal and finding of the semantic and hidden relation between keywords and documents. The LSA technique has also been used to uncover the correlation between audio-visual modalities for talking-face detection [74]. The learning correlation using LSA consists of four steps: construction of a joint multimodal feature space, normalization, singular value decomposition and measuring semantic association [28]. The mathematical details can be found in Li et al. [74]. In [74], the authors demonstrated the superiority of LSA over the traditional correlation coefficient.

Canonical correlation analysis. Canonical correlation analysis (CCA) is another powerful statistical technique that can be used to find linear mapping that maximizes the cross-correlation between two feature sets. Given two feature sets I_i and I_j, the CCA is a set of two linear projections $\mathbf{A}$ and $\mathbf{B}$ that whiten I_i and I_j. $\mathbf{A}$ and $\mathbf{B}$ are called canonical correlation matrices. These matrices are constructed under the constraints that their cross-correlation becomes diagonal and maximally compact in the projected space. The computation details of $\mathbf{A}$ and $\mathbf{B}$ can be found in [117]. The first M vectors of $\mathbf{A}$ and $\mathbf{B}$ are used to compute the synchrony score $\mathrm{CCA}(I_i, I_j)$ between two modalities I_i and I_j as:

$$\mathrm{CCA}(I_i, I_j) = \frac{1}{M}\sum_{m=1}^{M} |\mathrm{corr}(a_m^{\mathrm{T}} I_i, b_m^{\mathrm{T}} I_j)| \tag{11}$$

where a_m^{T} and b_m^{T} are elements of $\mathbf{A}$ and $\mathbf{B}$.

It is important to note that finding the canonical correlations and maximizing the mutual information between the sets are considered equivalent if the underlying distributions are elliptically symmetric [28].

The canonical correlation analysis for computing the synchrony score between modalities has been explored by few researchers. For instance, Chetty and Wagner [28] used the CCA score between audio and video modalities for

biometric person authentication. In this work, the authors also used LSA and achieved about 42% overall improvement in error rate with CCA and 61% improvement with LSA. In another work, Bredin and Chollet [20] also demonstrated the utility of considering CCA for audio–video based talking-face identity verification. Similarly, Slaney and Covell [117] used CCA for talking-face detection.

Cross-modal factor analysis. The weakness of the LSA method lies in its inability to distinguish features from different modalities in the joint space. To overcome this weakness, Li et al. [72] proposed the cross-modal factor analysis (CFA) method, in which, the features from different modalities are treated as two subsets and the semantic patterns between these two subsets are discovered. The method works as follows. Let the two subsets of features be I_i and I_j. The objective is to find the orthogonal transformation matrices A and B that can minimize the expression:

$$\left\| I_i A - I_j B \right\|_F^2 \tag{12}$$

where $A^T A$ and $B^T B$ are unit matrices. F denotes the Frobenius norm and is calculated for the matrix M as:

$$\|M\|_F = \left(\sum_x \sum_y |m_{xy}|^2 \right)^{1/2} \tag{13}$$

By solving the above equation for optimal transformation matrices A and B, the transformed version of I_i and I_j can be calculated as follows:

$$\tilde{I}_i = I_i A, \tilde{I}_j = I_j B \tag{14}$$

The optimized vectors $\tilde{I}_i$ and $\tilde{I}_j$ represent the coupled relationships between the two feature subsets I_i and I_j.

Note that, unlike CCA, the CFA provides a feature selection capability in addition to feature dimension reduction and noise removal. These advantages make CFA a promising tool for many multimedia analysis tasks. The authors in [72] have shown that although all three methods (LSA, CCA and CFA) achieved significant dimensionality reduction, the CFA gave the best results for talking head analysis. The CFA method achieved 91% detection accuracy as compared to the LSA (66.1%) and the CCA (73.9%).

All the methods described above have computed the correlation between the features extracted from different modalities. In the following, we describe the methods that have been used to compute the correlation at the semantic (or decision) level.

Causal link analysis. The events that happen in an environment are often correlated. For instance, the events of "elevator pinging", "elevator door opening", "people coming out of elevator" usually occur at relative times one after another. This temporal relationship between events has been utilized by Stauffer [123] for detecting events in a surveillance environment. This kind of analysis of the events has been called the casual link analysis.

Assuming that the two events were linked, the likelihood $p(c_i, c_j, \delta t_{ij} | \gamma_{i,j} = 1)$ of a pair (c_i, c_j) of events and their relative times δt_{ij} is estimated. Note that, δt_{ij} is the time difference between the absolute times of event i and event j. The term $\gamma_{i,j} = 1$ indicates that the occurrence of the first event c_i is directly responsible for the occurrence of the second event c_j. Once the estimates of the posterior likelihood of $\gamma_{i,j} = 1$ for all i, j pairs of events have been computed, an optimal chaining hypothesis is iteratively determined. The authors have demonstrated that the casual link analysis significantly helps in the overall accuracy of event detection in an audio–video surveillance environment.

Causal strength. The causal strength is a measure of the cause due to which two variables may co-vary with each other. Wu et al. [142] have preferred to use the causal strength between the context/content and the semantic ontology as described in Sect. 3.2.4. Here we describe how the causal strength is computed by adopting a probabilistic model. Let u and d be chance and decision variables, respectively. The chance variables imply effects (e.g. wearing warm jacket and drinking hot coffee) and the decision variables denote causes (e.g. cold weather). The causal strength $\mathrm{CS}_{u|d}$ is computed by using Eq. 15.

$$\mathrm{CS}_{u|d} = \frac{P(u|d, \xi) - P(u, \xi)}{1 - P(u, \xi)} \tag{15}$$

In the above equation, ξ refers to the state of the world, e.g. indoor or outdoor environment in the above mentioned example; and the terms $P(u|d, \xi)$ and $P(u, \xi)$ are the conditional probability assuming that d and ξ are independent of each other.

The authors have shown that the usage of causal strength provides not only improved accuracy of photo annotation, but also better capability of assessing the annotation quality.

Agreement coefficient. Atrey et al. [8] have used the correlation among streams at the intra-media stream and inter-media stream levels. At the intra-media stream level, they used the traditional correlation coefficient; however, at the inter-media stream level, they introduced the notion of the decision level "agreement coefficient". The agreement coefficient among streams has been computed based on how concurring or contradictory the evidence is that they provide. Intuitively, the higher the agreement among the streams, the more confidence one would have in the global decision, and vice versa [115].

The authors have modeled the agreement coefficient in the context of event detection in a multimedia surveillance scenario. The agreement coefficient $\gamma_{i,j}^k(t)$ between the

media streams I_i and I_j detects the kth event at time instant t, by iteratively averaging the past agreement coefficients with the current observation. Precisely, $\gamma^k_{i,j}(t)$ is computed as:

$$\gamma^k_{i,j}(t) = (1 - 2 \times |p_{i,k}(t) - p_{j,k}(t)|) + \gamma^k_{i,j}(t-1) \quad (16)$$

where, $p_{i,k}(t)$ and $p_{j,k}(t)$ are the individual probabilities of the occurrence of kth event based on the media streams I_i and I_j, respectively, at time $t \geq 1$; and $\gamma^k_{i,j}(0) = 1 - 2 \times |p_{i,k}(0) - p_{j,k}(0)|$. These probabilities represent decisions about the events. Exactly the same probabilities would imply full agreement ($\gamma^k_{i,j} = 1$) while detecting the kth event whereas totally dissimilar probabilities would mean that the two streams fully contradict each other ($\gamma^k_{i,j} = -1$). The authors have shown that the usage of agreement coefficient resulted in better overall event detection accuracy in a surveillance scenario.

Independence. In should be noted that, in addition to using the correlation among modalities, the independent modalities can also be very useful in some cases to obtain a better decision. Let us consider the case of a multimodal dialog system [49, 96, 100]. In such systems, multiple modalities such as gesture and speech are used as a means of interaction. It is sometimes very hard to fuse these modalities at the feature level due to a lack of direct correspondence between their features and different temporal alignment. However, each modality can complement each other in obtaining a decision about the intended interaction event. To this regard, each modality can be processed separately in parallel to derive individual decisions and later fuse these individual decisions at a semantic level to obtain the final decision [96]. Similarly, other cases of independence among modalities are also possible. For instance, environment context, device context, network context, task context and so forth may provide complementary information to the fusion process, thereby making the overall analysis tasks more robust and accurate.

4.1.2 Confidence level of different modalities

Different modalities may have varying capabilities of accomplishing a multimedia analysis task. For example, in good lighting condition, the video analysis may be more useful in detecting human than the audio analysis; while in a dark environment, the audio analysis could be more handy. Therefore, in the fusion process, it is important to assign the appropriate confidence level to the participating streams [115]. The confidence in a stream is usually expressed by assigning appropriate weight to it.

Many fusion methods such as the linear weight fusion and the Bayesian inference do have a notion of specifying the weights to different modalities. However, the main question that remains to be answered is how to determine the weights of different modalities. These weights can vary based on several factors such as the context and the task performed. Therefore, the weight should be dynamically adjusted in order to obtain optimal fusion results.

While performing multimodal fusion, several researchers have adopted the strategy of weighting different modalities. However, many of them either have considered equal weights [83, 152] or have not elaborated the issue of weight determination [55, 67, 136], and have left it to the users to decide [40].

Other works, such as Neti et al. [87], Iyenger et al. [57], Tatbul et al. [126], Hsu and Chang [53] and Atrey et al. [8] have used pre-computed weights in the fusion process. The weights of different streams have usually been determined based on their past accuracy or any prior knowledge. The computation of past accuracy requires a significant amount of training and testing. However, since the process of computing the accuracy has to be performed in advance, the confidence value or the weight determined based on such accuracy value is considered "static" during the fusion process. It is obvious that a static value of confidence of a modality does not reflect its true current value especially under the changing context. On the other hand, determining the confidence level for each stream, based on its past accuracy, is difficult. This is because the system may provide dissimilar accuracies for various tasks under different contexts. Pre-computation of accuracies of all the streams for various detection tasks under varying contexts requires a significant amount of training and testing, which is often tedious and time consuming. Therefore, a mechanism that can determine the confidence levels of different modalities "on the fly" without pre-computation, needs to be explored.

In contrast to the above methods that used the static confidence, some efforts (e.g. Tavakoli et al. [127], Atrey et al. [10] have also been performed towards the dynamic computation of the confidence levels of different modalities. Tavakoli et al. [127] have used spatial and temporal information in clusters in order to determine a confidence level of sensors. The spatial information indicated that more sensors are covering a specific area; hence a higher confidence is assigned to the observation obtained from that area. The temporal information is obtained in the form of the sensors detecting the target consecutively for a number of time slots. If a target is consecutively detected, it was assumed that the sensors are reporting correctly. This method is more suited to the environment where the sensors' location changes over time. In a fixed sensor setting, the confidence value will likely remain constant.

Recently, Atrey et al. [10] have also presented a method to dynamically compute the confidence level of a media stream based on its agreement coefficient with a trusted stream. The trusted stream is the one that has the confidence level above a certain threshold. The agreement coefficient

between any two streams will be high when the similar decisions are obtained based on them, and vice versa. In this work, the authors have adopted the following idea. Let one follow a trusted news bulletin. He/she also starts by following an arbitrary news bulletin and compares the news content provided on both the news bulletins. Over a period of time, his/her confidence in the arbitrary bulletin will also grow if the news content of both the bulletins have been found similar, and vice versa. The authors have demonstrated that the confidence level of different media streams computed using this method when used in the fusion process provides the event detection results comparable to what is obtained using pre-computed confidence. The drawback with this method is that the assumption of having at least one trusted stream might not always be realistic.

The above discussion shows that, although there have been some attempts to address the issue of dynamic weight adjustment, this is still an open research problem, which is essential for the overall fusion process. A summarization of the representative works related to the computation of confidence provided in Table 6.

4.1.3 Contextual information

The context is accessory information that greatly influences the performance of a fusion process. For example, time and location information significantly improves the accuracy of automatic photo classification [142]. Also, the light conditions may help in selecting the right set of sensors for detecting events in a surveillance environment.

Some of the earlier works, which have emphasized the importance of using the contextual information in the fusion process, include Brmond and Thonnat [21], Teriyan and Puuronen [129], Teissier et al. [128] and Westerveld [139]. Later, many other researchers such as Sridharan et al. [122], Wang and Kankanhalli [135], Pfleger [100] and Atrey et al. [8] have demonstrated the advantages of using context in the fusion process.

Two research issues related to the context are (1) what are the different forms of contextual information and how the contextual information is determined? and (2) how it is used in the fusion process? In the following, we discuss how these two issues have been addressed by the researchers.

The context has been represented in different forms for different multimedia analysis tasks. For example, for the image classification task, the context could be time, location and camera parameters [142], while for the multimedia music selection task, the mood of the requester could be considered as context. We identify two types of contextual information that have been often considered. These are environmental context and the situational context. The environmental context consists of time, the sensor's location or geographical location, weather conditions, etc. For example, if it is a dark environment, audio and IR sensor information should preferably be fused to detect a person [8]. The situational context could be in the form of identity, mood, and capability of a person, etc. For example, if the person's mood is happy, a smart mirror should select and play a romantic song when s/he enters into a smart house [50].

The contextual information can be determined by explicitly processing the sensor data, e.g. a mood detection algorithm can be applied on the video data to determine the mood of a person. On the other hand, it can also be learned through other mechanisms such as the time from a system clock, location from a GPS device, and the sensors' geometry and location as a priori information from the system designer.

To integrate the contextual information in the fusion process, most researchers such as Westerveld [139], Jasinschi et al. [62], Wang and Kankanhalli [135], Pfleger [100], Wu et al. [142], Atrey et al. [8] have adopted a rule-based scheme. This scheme is very straight forward as it follows the "if–then–else" strategy. For example, *if* it is day time, *then* the video cameras would be assigned a greater weight than the audio sensors in the fusion process for detecting the event of a "human walking in the

Table 6 A list of the representative works related to the usage of confidence level in the fusion process

The mode of computation	The works	Multimedia analysis task	The confidence is determined based on
Static	Neti et al. [87]	Speaker recognition and speech event detection	The past accuracy
	Iyenger et al. [57]	Monologue detection	
	Tatbul et al. [126]	Military smart uniform	
	Hsu and Chang [53]	News video analysis	
	Atrey et al. [8]	Event detection for surveillance	
Dynamic	Tavakoli et al. [127]	Event detection in undersea sensor networks	The spatial and the temporal information of sensors' observations
	Atrey et al. [10]	Event detection for surveillance	The agreement/disagreement between different streams

garden", *else* otherwise. We describe some of these works in the following. In [139], the author has integrated image features (content) and the textual information that comes with an image (context) at the semantic level. Similar to this work, Jasinschi et al. [62] have presented a layered probabilistic framework that integrates the multimedia content and context information. Within each layer, the representation of content and context is based on Bayesian networks, and hierarchical priors that provide the connection between the two layers. The authors have applied the framework for an end-to-end system called the video scout that selects, indexes, and stores TV program segments based on topic classification. In the context of dialog systems, Pfleger [100] has presented a multimedia fusion scheme for detecting the user actions and events. While detecting these input events, the user's 'local turn context' has been considered. This local turn context comprises all previously recognized input events and the dialog states that both belong to the same user's turn. Wu et al. [142] have used the context information (in form of the time and location) for photo annotation. They have adopted a Bayesian network fusion approach in which the context has been used to govern the transitions between nodes. Wang and Kankanhalli [135] and Atrey et al. [8] have used the context in the form of the environment and the sensor information. The environment information consisted of the geometry of the space under surveillance while the sensory information was related to their location and orientation.

While the works described above have used the context in a static manner, Sridharan et al. [122] have provided a computational model of context evolution. The proposed model represents the context using semantic-nets. The context has been defined as the union of semantic-nets, each of which can specify a fact about the environment. The interrelationships among the various aspects (e.g. the user, the environment, the allowable interactions, etc.) of the system are used to define the overall system context. The evolution of context has been modeled using a leaky bucket algorithm that has been widely used for traffic control in a network.

The representative works related to the use of contextual information in the fusion process have been summarized in Table 7. Although the rule-based strategy of integrating the context in the fusion process is appealing, the number of rules largely increases in varying context in a real world scenario. Therefore, other strategies for context determination and its integration in multimodal fusion remain to be explored in future.

4.2 Issue related to when to fuse

Different modalities are usually captured in different formats and at different rates. Therefore, they need to be synchronized before fusion takes place [95]. As the fusion can be performed at the feature as well as the decision level, the issue of synchronization is also considered at these two levels. In the feature level synchronization, the features obtained from different but closely coupled modalities captured during the same time period are combined together [28]. On the other hand, the decision level synchronization needs to determine the designated points along the timeline at which the decisions should be fused. However, in both the levels of fusion, the problem of synchronization arises in different forms. In the following, we elaborate on these problems and also describe some works which have addressed them.

The feature level synchronization has been illustrated in Fig. 9a. Assuming that the raw data from the two different types of modalities (modality 1 and modality 2, in the figure) are obtained at the same time $t = 1$. The feature extraction from these modalities can be from different time periods (e.g. 2 and 1.5 time units for modality 1 and modality 2, respectively in Fig. 9a). Due to the different time periods of the data processing and feature extraction, when these two features should be combined, remains an issue. To resolve this issue, a simple strategy could be to fuse the features at regular intervals [8]. Although this strategy may not be the best, it is computationally less expensive. An alternative strategy could be to combine all

Table 7 The representative works related to the use of contextual information in the fusion process

Contextual information	The works	Multimedia analysis task
Textual information with an image	Westerveld [139]	Image retrieval
Signature, pattern, or underlying structure in audio, video and transcript	Jasinschi et al. [62]	TV program segmentation
Environment and sensor information	Wang and Kankanhalli [135], Atrey et al. [8]	Event detection for surveillance
Word nets	Sridharan et al. [122]	Multimedia visualization and annotation
Past input events and the dialog state	Pfleger [100]	Detecting the user intention in multimodal dialog systems
Time, location and camera parameters	Wu et al. [142]	Photo annotation

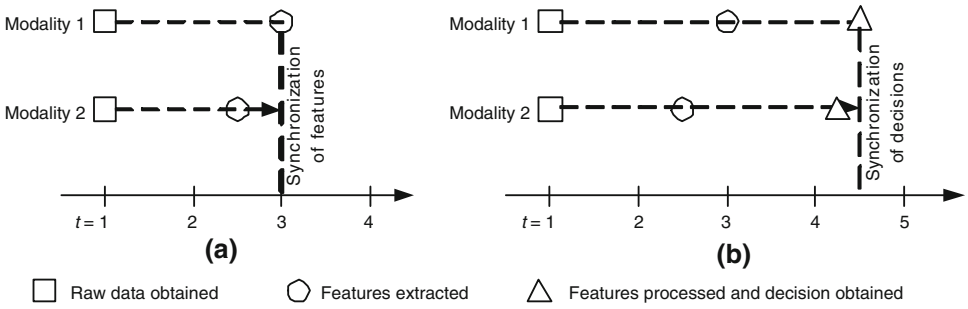

Fig. 9 Illustration of the synchronization between two modalities at the **a** feature level, **b** decision level

the features at the time instant they are available (e.g. at $t = 3$ in Fig. 9a).

An illustration of the synchronization at the decision level is provided in Fig. 9b. In contrast to feature level synchronization, where only the feature extraction time impact asynchrony between the modalities, the additional time of obtaining decisions based on the extracted features further affect it. For example, as shown in Fig. 9b, the time taken in obtaining the decisions could be 1.5 and 1.75 time units for modality 1 and modality 2, respectively. However, similar to feature level synchronization, in this case also, the decisions are fused using various strategies discussed earlier (e.g. at the time instant all the decisions are available, $t = 4$ in Fig. 9b). Exploring the best strategy is an issue that can be considered in future.

Another important synchronization issue is to determine the amount of raw data needed from different modalities for accomplishing a task. To mark the start and end of a task (e.g. event detection) over a timeline, there is a need to obtain and process the data streams at certain time intervals. For example, from a video stream of 24 frames/s, 2-s data (48 frames) could be sufficient to determine a human walking event (by computing the blob displacement in a sequence of images); however, the same event (sound of footsteps) could be detected using one second of audio data of 44 kHz. This time period, which is basically the minimum amount of time to accomplish a task, could be different for different tasks when accomplished using various modalities. Ideally, it should be as small as possible since a smaller value allows task accomplishment at a finer granularity in time. In other words, the minimum time period for a specific task should be just large enough to capture the data to accomplish it. Determining the minimum time period to accomplish different tasks is a research issue that needs to be explored in future.

In multimedia fusion literature, the issue of synchronization has not been widely addressed. This is because many researchers focused on the accuracy aspect of the analysis tasks and performed experiments in an offline manner. In the offline mode, the synchorization has often been manually performed by aligning the modalities along a timeline. The researchers who performed analysis tasks in real time have usually adopted simple strategies such as synchronization at a regular interval. However, having a regular interval may not be optimal and may not lead to the accomplishment of the task with the highest accuracy. Therefore, the issue of synchronization still remains to be explored.

In the following, we discuss some representative works that have focused on synchronization issue in one way or the other. In the area of audio-visual speech processing, several researchers have computed the audio-visual synchrony. These works include Hershey and Movellan [47], Slaney and Covell [117], Iyengar et al. [57], Nock et al. [91] and Bredin and Chollet [19]. In these works, the audio-visual synchrony has been used as a measure of correlation that can be perceived as synchronization at the feature level.

The problem of synchronization at the decision level, which is more difficult than the feature level synchronization, has been addressed by few researchers including Holzapfel et al. [49], Atrey et al. [8], and [Xu and Chua [149]. Holzapfel et al. [49] have aligned the decisions obtained from the processing of gesture and speech modalities. To identify the instances at which these two decisions are to be along the timeline, the authors computed temporal correlation between the two modalities. Unlike Holzapfel et al. [49], Atrey et al. [8] adopted a simple strategy to combine the decisions at regular intervals. These decisions were obtained from audio and video event detectors. The authors have empirically found that the time interval of one second was optimal in improving the overall accuracy of event detection.

The issue of time synchronization has also been widely addressed in news and sports video analysis. Satoh et al. [114] adopted a multimodal approach for face identification and naming in news video by aligning the text, audio and video modalities. The authors proposed to detect face sequences from images and extract the name candidates from the transcripts. The transcripts were obtained from the audio tracks using speech recognition technique. Moreover, video captions were also processed to extract the name titles. Based on the audio-generated transcript and the video captions, the corresponding faces in the video were aligned.

In the domain of sports event analysis, to determine the time when the event occurred in the broadcasted sports video, Babaguchi et al. [13] used the textual overlays that usually appear in the sports video. Similar approach was used by Xu and Chua [149]. In their work, the authors have used text modality in addition to the audio and video. The authors observed a significant asynchronism that resulted from different time granularities of audio–video and text analysis. While the audio–video frames were available at a regular interval of seconds, the text availability was very slow (approximately every minute). This was because, the human operator usually enters texts for live matches which takes a few minutes to become available for automatic analysis. The authors have performed synchronization between text and audio–video modalities by using alignment. This alignment was performed by maximizing the number of matches between text events and audio–video events. The text and audio–video events are considered matched when they are both within a temporal range, occur in the same sequential order, and the audio–video event adapts to the modeling of the text event. For example, an offense followed by a break conforms to goal's event modeling. Although the above mentioned synchronization method works well, it cannot be generalized since it is domain-oriented and highly specific to the user-defined rules. Note that, while Babaguchi et al. [13] and Xu and Chua [149] attempted to synchronize video based on the time extracted from the time overlays in the video and web-casted text, respectively; Xu et al. [147] and [148] adopted a different approach. In their work, the authors extracted the timing information from the broadcasted video by detecting the video event boundaries. The authors observed that as the webcasted text is usually not available in the broadcasted text, the time recognition from the broadcast sports video is a better choice to perform the alignment of the text and video.

A summarization of the above described works has been provided in Table 8.

4.3 Issue related to what to fuse

In the literature, the issue of what to fuse has been addressed at two different levels: modality selection and feature vector reduction. Modality selection refers to choosing different types of modalities. For example, one can select and fuse two video camera and one microphone data to determine the presence of a person. On the other hand, the fusion of features usually results into a large feature vector, which becomes a bottleneck for a particular analysis task. This is known as the curse of dimensionality. To overcome this problem, different data reduction techniques are applied to reduce the feature vector. In the following, we discuss various works that addressed the modality selection and feature vector reduction issues.

4.3.1 Modality selection

The modality selection problem is similar to the sensor stream selection problem that has often been considered as an optimization problem, in which, the best set of sensors is obtained satisfying some cost constraints. Some of the fundamental works on the sensor stream selection in the context of discrete-event systems and failure diagnosis include Oshman [94], Debouk et al. [37] and Jiang et al. [64]. Similarly, in the context of wireless sensor networks, the optimal media stream selection has been studied by Pahalawatta et al. [98], Lam et al. [70], and Isler and Bajcsy [56]. The details of these works are omitted in the interest of brevity.

In the context of multimedia analysis, the problem of optimal modality selection has been targeted by Wu et al. [143], Kankanhalli et al. [68], and Atrey et al. [9]. In [143], the authors proposed a two-step optimal fusion approach. In the first step, they find statistically independent modalities from raw features. Then, the second step involves super-kernel fusion to determine the optimal combination of individual modalities. The authors have provided a tradeoff between modality independence and the curse of

Table 8 A list of representative works that have addressed synchronization problem

Level of fusion	The work	Multimedia analysis task
Feature	Hershey and Movellan [47], Slaney and Covell [117], Nock et al. [91]	Speech recognition and speaker localization
	Iyengar et al. [57]	Monologue detection
	Bredin and Chollet [19]	Biometric-based person identification
Decision	Holzapfel et al. [49]	Dialog understanding
	Atrey et al. [8]	Event detection for surveillance
	Satoh et al. [114]	News video analysis
	Babaguchi et al. [13], Xu and Chua [149], Xu et al. [147, 148]	Sports video analysis

dimensionality. Their idea of selecting optimal modalities is as follows. When the number of modalities is one, all the feature components were treated as a one-vector representation, suffering from the curse of dimensionality. On the other hand, the large number of modalities reduces the curse of dimensionality, but the inter modality correlation is increased. An optimal value of modalities would tend to balance between the curse of dimensionality and the inter modality correlation. The authors have demonstrated the utility of their scheme for image classification and video concept detection. Kankanhalli et al. [68] have also presented an Experiential Sampling based method to find the optimal subset of streams in multimedia systems. Their method is the extension of the work by Debouk et al. [37], in the context of multimedia. This method may have a high cost of computing the optimal subset as it requires the minimum expected number of tests to be performed in order to determine the optimal subset.

Recently, Atrey et al. [9] have presented a dynamic programming based method to select the optimal subset of media streams. This method provides a threefold tradeoff between the extent to which the multimedia analysis task is accomplished by the selected subset of media streams, the overall confidence in this subset, and the cost of using this subset. In addition, their method also provides flexibility to the system designer to choose the next best sensor if the best sensor is not available. They have demonstrated the utility of the proposed method for event detection in an audio–video surveillance scenario.

From the above discussion, it can be observed that only a few attempts (Wu et al. [143], Kankanhalli et al. [68], and Atrey et al. [9]) have been made to select the best (or optimal) subset of modalities for multimodal fusion. However, these methods have their own limitations and drawbacks. Moreover, they do not consider the different contexts under which the modalities may be selected. Therefore, a lot more can be done in this aspect of multimodal fusion. The methods for optimal subset modality selection described above are summarized in Table 9.

4.3.2 Feature vector reduction

It is important to mention that, besides selecting the optimal set of the modalities, the issue "what to fuse" for a particular multimedia analysis task involves the reduction of feature vector. The fusion of features that are obtained from different modalities usually result into a large feature vector, which becomes a bottleneck when processed to accomplish any multimedia analysis task. To handle such situations, researchers have used various data reduction techniques. Most commonly used techniques are principle component analysis (PCA), singular vector decomposition (SVD) and linear discriminant analysis (LDA).

PCA is used to project higher dimensional data into lower dimensional space while preserving as much information as possible. The projection that minimizes the squared error in reconstructing original data is chosen to represent the reduced set of features. The PCA technique often does not perform well when the dimensionality of the feature set is very large. This limitation is overcome by SVD technique, which is used to determine the eigen vectors that most represent the input feature set. While PCA and SVD are unsupervised techniques, LDA works in supervised mode. LDA is used for determining the linear combination of features, which is not only a reduced set of features but it is also used for classification. The readers may refer to [134] for further details about these feature dimensionality reduction methods.

In multimodal fusion domain, many researchers have used these methods for feature vector dimension reduction. Some representative works are: Guironnet et al. [44] used PCA for video classification, Chetty and Wagner [28] utilized SVD for biometric person authentication, and Potamianos et al. [105] adopted LDA for speech recognition.

4.3.3 Other considerations

There are other situations when the issue of "what to fuse" needs special consideration. For instance, dealing with unlabeled data in fusion [130] and handling noisy positive data for fusion [54].

There are three approaches used for learning with unlabeled data: semi-supervised learning, transductive learning and active learning [155]. Semi-supervised learning methods automatically exploit unlabeled data to help estimate the data distribution in order to improve learning performance. Transductive learning is different from semi-supervised learning in that it selects the

Table 9 A summary of approaches used for optimal modality subset selection

The work	The optimality criteria	Drawback
Wu et al. [143]	Curse of dimensionality versus inter modality correlation	The gain from a modality is overlooked
Kankanhalli et al. [68]	Information gain versus cost (time)	Cost of computing the optimal subset could be high
Atrey et al. [9]	Gain versus Confidence level versus processing cost	The issue of how frequently the optimal subset should be recomputed needs a formalization

unlabeled data from test data set. On the other hand, in active learning methods, the learning algorithm selects the unlabeled example and actively query the user/teacher for labels. Here the learning algorithm is supposed to be good enough to choose the least number of examples to learn a concept, otherwise there is a risk of including the unimportant and irrelevant examples.

Another important issue that needs to be resolved is how to reduce outliers or noise in the input data for fusion. In the multimodal fusion process, noisy data usually results into reduced classification accuracy and increased training time and size of the classifier. There are various solutions to deal with the noisy data. For instance, to employ some noise filter mechanisms to smooth the noisy data or to apply an appropriate sampling technique to differentiate the noisy data from the input data before the fusion takes place [137].

5 Benchmark datasets and evaluation

5.1 Datasets

Many multimodal fusion applications use several publicly available datasets. For example, the small $2k$ image datasets in Corel Image CDs. It contains representative images of fourteen categories that includes architecture, bears, clouds, elephants, fabrics, fireworks, flowers, food, landscape, people, textures, tigers, tools, and waves. Different features such as color and texture can be extracted from this $2k$ image dataset and be used for fusion as shown in [143].

Among the video-based fusion research, the most popular are the well-known TRECVID datasets [2] that are available in different versions since 2001. A quick view of the high-level feature extraction from these datasets can be found in [118]. Depending on their release, these datasets contain data files about broadcast news video, sound and vision video, BBC rushes video, BBC rushes video, London Gatwick surveillance video, and test dataset annotations for surveillance event detection. Features from visual, audio and caption tracks in TRACVID datasets are extracted and used in fusion for various multimedia analysis tasks, such as video shot retrieval [83], semantic video analysis [121], news video story segmentation [52], video concept detection [58, 143] and so on.

The fusion literature related to biometric identification and verification make ongoing efforts to build multimodal biometric databases. For example, BANCA [14] that contains face and speech modalities; XM2VTS [82] that contains synchronized video and speech data; BIOMET [42] that contains face, speech, fingerprint, hand and signature modalities; MYCT [93] that contains 10-print fingerprint and signature modalities and several others as mentioned in [104].

Another popular dataset standardization effort has been the agenda of performance evaluation of tracking and surveillance (PETS) community [1]. Several researchers have used PETS datasets for multimodal analysis tasks, for example, object tracking [154].

Although there are several available datasets that can be used for various analysis tasks, there lacks any standardization effort for a common dataset for multimodal fusion research.

5.2 Evaluation measures

Several evaluation metrics are usually used to measure the performance of the fusion-based multimedia analysis tasks. For example, NIST average precision metric is used to determine the accuracy of semantic concept detection at the video shot level [58, 121, 143]. For news video story segmentation, the precision and recall metrics are widely used [52]. Precision and recall measure are also commonly used for image retrieval [55]. Similarly, for the video shot retrieval some researchers use mean average precision [83]. While performing image categorization, the accuracy of the classification is measured in terms of image category detection rate [156].

To measure the performance of tracking related analysis tasks, the dominating evaluation metrics include mean distance from track, detection rate, false positive rate, recall and precision [131]. Similarly, for speaker position estimation researchers have measured tracking error and calculated average distance between true and estimated position of speaker [125]. In [154], the authors calculated variance of motion direction and variance of compactness to calculate the accuracy of object tracking.

Recently, Hossain et al. [51] presented a multi-criteria evaluation metric to determine the quality of information obtained based on multimodal fusion. The evaluation metric includes certainty, accuracy and timeliness. The authors showed its applicability in the domain of multimedia monitoring.

In human computer interaction, fusion is used mostly to identify multimodal commands or input interactions of human such as gestures, speech etc. Therefore, metrics such as speech recognition accuracy and gesture recognition accuracy are used to measure the accuracy of these tasks [49].

Furthermore, to evaluate the fusion result for biometric verification, false acceptance rate (FAR) and false rejection rate (FRR) are used to identify the types of errors [104]. The FAR and FRR are often used to present the half total error rate (HTER), which is a measure to assess the quality of a biometric verification system [17].

Overall, we observed that the researchers have used different evaulation criteria for different analysis tasks.

However a common evaluation framework for multimodal fusion is yet to be achieved.

6 Conclusions and future research directions

We have surveyed the state-of-the-art research related to multimodal fusion and commented on these works from the perspective of the usage of different modalities, the levels of fusion, and the methods of fusion. We have further provided a discussion to summarize our observation based on the reviewed literature, which can be useful for the readers to have an understanding of the appropriate fusion methodology and the level of fusion. Some distinctive issues (e.g. correlation, confidence) that influence the fusion process are also elaborated in greater detail.

Despite that a significant number of multimedia analysis tasks have been successfully performed using a variety of fusion methods, there are several areas of investigation that may be explored in the future. We have identified some of them as follows:

1. Multimedia researchers have mostly used the audio, video and the text modalities for various multimedia analysis tasks. However, the integration of some new modalities such as RFID for person identification, haptics for dialog systems, etc. can be explored further.
2. The appropriate synchronization of the different modalities is still a big research problem for multimodal fusion researchers. Specifically, when and how much data should be processed from different modalities to accomplish a multimedia analysis task, is an issue that has not yet been explored exhaustively.
3. The problem of the optimal weight assignment to the different modalities under a varying context is an open problem. Since we usually have different confidence levels in the different modalities for accomplishing various analysis tasks, the problem of dynamic computation of the confidence information for the different streams for various tasks, becomes challenging and worth researching in future.
4. How to integrate context in the fusion process? This question can be answered by thinking beyond the "if–then–else" strategy. There is a need to formalize the concept of context. How may the changing context influence the fusion process? What model would be most suitable to simulate the varying nature of context? These questions require greater attention from multimedia researchers.
5. The feature level correlation among different modalities has been utilized in an effective way. However, it has been observed that correlation at the semantic level (decision level) has not been fully explored, although some initial attempts have been reported.
6. The optimal modality selection for fusion is emerging as an important research issue. From the available set, which modalities should be fused to accomplish a task at a particular time instant? The utility of these modalities could be changed with the varying context. Moreover, the optimality of modality selection can be determined based on various constraints such as the extent to which the task is accomplished, the confidence with which the task is accomplished, and the cost of using the modalities for performing the task. As the optimal subset changes over time, how frequently it should be computed so that the cost of re-computation can be reduced to meet the timeliness, is an open problem for multimedia researchers to consider.
7. Last but not least, there are various evaluation metrics that are used to measure the performance of different multimedia analysis tasks. However, it would be interesting to work on a common evaluation framework that can be used by multimodal fusion community.

Multimodal fusion for multimedia analysis is a promising research area. This survey has covered the existing works in this domain and identified several relevant issues that deserve further investigation.

Acknowledgments The authors would like to thank the editor and the anonymous reviewers for their valuable comments in improving the content of this paper. This work is partially supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

1. PETS: Performance evaluation of tracking and surveillance (Last access date 31 August 2009). http://www.cvg.rdg.ac.uk/slides/pets.html
2. TRECVID data availability (Last access date 02 September 2009). http://www-nlpir.nist.gov/projects/trecvid/trecvid.data.html
3. Adams, W., Iyengar, G., Lin, C., Naphade, M., Neti, C., Nock, H., Smith, J.: Semantic indexing of multimedia content using visual, audio, and text cues. EURASIP J. Appl. Signal Process. **2003**(2), 170–185 (2003)
4. Aguilar, J.F., Garcia, J.O., Romero, D.G., Rodriguez, J.G.: A comparative evaluation of fusion strategies for multimodal biometric verification. In: International Conference on Video-Based Biometrie Person Authentication, pp. 830–837. Guildford (2003)
5. Aleksic, P.S., Katsaggelos, A.K.: Audio-visual biometrics. Proc. IEEE **94**(11), 2025–2044 (2006)
6. Andrieu, C., Doucet, A., Singh, S., Tadic, V.: Particle methods for change detection, system identification, and control. Proc. IEEE **92**(3), 423–438 (2004)
7. Argillander, J., Iyengar, G., Nock, H.: Semantic annotation of multimedia using maximum entropy models. In: International

Conference on Accoustic, Speech and Signal Processing, pp. II–153–156. Philadelphia (2005)
8. Atrey, P.K., Kankanhalli, M.S., Jain, R.: Information assimilation framework for event detection in multimedia surveillance systems. Springer/ACM Multimed. Syst. J. **12**(3), 239–253 (2006)
9. Atrey, P.K., Kankanhalli, M.S., Oommen, J.B.: Goal-oriented optimal subset selection of correlated multimedia streams. ACM Trans. Multimedia Comput. Commun. Appl. **3**(1), 2 (2007)
10. Atrey, P.K., Kankanhalli, M.S., El Saddik, A.: Confidence building among correlated streams in multimedia surveillance systems. In: International Conference on Multimedia Modeling, pp. 155–164. Singapore (2007)
11. Ayache, S., Quénot, G., Gensel, J.: Classifier fusion for svm-based multimedia semantic indexing. In: The 29th European Conference on Information Retrieval Research, pp. 494–504. Rome (2007)
12. Babaguchi, N., Kawai, Y., Kitahashi, T.: Event based indexing of broadcasted sports video by intermodal collaboration. IEEE Trans. Multimed. **4**, 68–75 (2002)
13. Babaguchi, N., Kawai, Y., Ogura, T., Kitahashi, T.: Personalized abstraction of broadcasted american football video by highlight selection. IEEE Trans. Multimed. **6**(4), 575–586 (2004)
14. Bailly-Bailliére, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., Matas, J., Messer, K., Popovici, V., Porée, F., Ruíz, B., Thiran, J.P.: The BANCA database and evaluation protocol. In: International Conference on Audio-and Video-Based Biometrie Person Authentication, pp. 625–638. Guildford (2003)
15. Beal, M.J., Jojic, N., Attias, H.: A graphical model for audio-visual object tracking. IEEE Trans. Pattern Anal. Mach. Intell. **25**, 828– 836 (2003)
16. Bendjebbour, A., Delignon, Y., Fouque, L., Samson, V., Pieczynski, W.: Multisensor image segmentation using Dempster–Shafer fusion in markov fields context. IEEE Trans. Geosci. Remote Sens. **39**(8), 1789–1798 (2001)
17. Bengio, S.: Multimodal authentication using asynchronous hmms. In: The 4th International Conference Audio and Video Based Biometric Person Authentication, pp. 770–777. Guildford (2003)
18. Bengio, S., Marcel, C., Marcel, S., Mariethoz, J. Confidence measures for multimodal identity verification. Inf. Fusion **3**(4), 267–276 (2002)
19. Bredin, H., Chollet, G.: Audio-visual speech synchrony measure for talking-face identity verification. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 233–236. Paris (2007)
20. Bredin, H., Chollet, G.: Audiovisual speech synchrony measure: application to biometrics. EURASIP J. Adv. Signal Process. 11 p. (2007). Article ID 70186
21. Brémond, F., Thonnat, M.: A context representation of surveillance systems. In: European Conference on Computer Vision. Orlando (1996)
22. Brooks, R.R., Iyengar, S.S.: Multi-sensor Fusion: Fundamentals and Applications with Software. Prentice Hall PTR, Upper Saddle River, NJ (1998)
23. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. **2**(2), 121–167 (1998)
24. Caruana, R., Munson, A., Niculescu-Mizil, A.: Getting the most out of ensemble selection. In: ACM International Conference on on Data Mining, pp. 828–833. Maryland (2006)
25. Chaisorn, L., Chua, T.S., Lee, C.H., Zhao, Y., Xu, H., Feng, H., Tian, Q.: A multi-modal approach to story segmentation for news video. World Wide Web **6**, 187–208 (2003)
26. Chang, S.F., Manmatha, R., Chua, T.S.: Combining text and audio-visual features in video indexing. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, pp. 1005–1008. IEEE Computer Society, Philadelphia (2005)
27. Chen, Q., Aickelin, U.: Anomaly detection using the dempster–shafer method. In: International Conference on Data Mining, pp. 232–240. Las Vegas (2006)
28. Chetty, G., Wagner, M.: Audio-visual multimodal fusion for biometric person authentication and liveness verification. In: NICTA-HCSNet Multimodal User Interaction Workshop, pp. 17–24. Sydney (2006)
29. Chieu, H.L., Lee, Y.K.: Query based event extraction along a timeline. In: International ACM Conference on Research and Development in Information Retrieval, pp. 425–432. Sheffield (2004)
30. Choudhury, T., Rehg, J.M., Pavlovic, V., Pentland, A.: Boosting and structure learning in dynamic bayesian networks for audio-visual speaker detection. In: The 16th International Conference on Pattern Recognition, vol. 3, pp. 789–794. Quebec (2002)
31. Chua, T.S., Chang, S.F., Chaisorn, L., Hsu, W.: Story boundary detection in large broadcast news video archives: techniques, experience and trends. In: ACM International Conference on Multimedia, pp. 656–659. New York, USA (2004)
32. Corradini, A., Mehta, M., Bernsen, N., Martin, J., Abrilian, S.: Multimodal input fusion in human–computer interaction. In: NATO-ASI Conference on Data Fusion for Situation Monitoring, Incident Detection, Alert and Response Management. Karlsruhe University, Germany (2003)
33. Crisan, D., Doucet, A.: A survey of convergence results on particle filtering methods for practitioners. IEEE Trans. Signal Process. **50**(3), 736–746 (2002)
34. Cutler, R., Davis, L.: Look who's talking: Speaker detection using video and audio correlation. In: IEEE International Conference on Multimedia and Expo, pp. 1589–1592. New York City (2000)
35. Darrell, T., Fisher III, J.W., Viola, P., Freeman, W.: Audio-visual segmentation and "the cocktail party effect". In: International Conference on Multimodal Interfaces. Bejing (2000)
36. Datcu, D., Rothkrantz, L.J.M.: Facial expression recognition with relevance vector machines. In: IEEE International Conference on Multimedia and Expo, pp. 193–196. Amsterdam, The Netherlands (2005)
37. Debouk, R., Lafortune, S., Teneketzis, D.: On an optimal problem in sensor selection. J. Discret. Event Dyn. Syst. Theory Appl. **12**, 417–445 (2002)
38. Ding, Y., Fan, G.: Segmental hidden markov models for view-based sport video analysis. In: International Workshop on Semantic Learning Applications in Multimedia. Minneapolis (2007)
39. Fisher-III, J., Darrell, T., Freeman, W., Viola, P.: Learning joint statistical models for audio-visual fusion and segregation. In: Advances in Neural Information Processing Systems, pp. 772–778. Denver (2000)
40. Foresti, G.L., Snidaro, L.: A distributed sensor network for video surveillance of outdoor environments. In: IEEE International Conference on Image Processing. Rochester (2002)
41. Gandetto, M., Marchesotti, L., Sciutto, S., Negroni, D., Regazzoni, C.S.: From multi-sensor surveillance towards smart interactive spaces. In: IEEE International Conference on Multimedia and Expo, pp. I:641–644. Baltimore (2003)
42. Garcia Salicetti, S., Beumier, C., Chollet, G., Dorizzi, B., les Jardins, J., Lunter, J., Ni, Y., Petrovska Delacretaz, D.: BIOMET: A multimodal person authentication database including face, voice, fingerprint, hand and signature modalities. In: International Conference on Audio-and Video-Based Biometrie Person Authentication, pp. 845–853. Guildford, UK (2003)
43. Gehrig, T., Nickel, K., Ekenel, H., Klee, U., McDonough, J.: Kalman filters for audio–video source localization. In: IEEE

Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 118– 121. Karlsruhe University, Germany (2005)
44. Guironnet, M., Pellerin, D., Rombaut, M.: Video classification based on low-level feature fusion model. In: The 13th European Signal Processing Conference. Antalya, Turkey (2005)
45. Hall, D.L., Llinas, J.: An introduction to multisensor fusion. In: Proceedings of the IEEE: Special Issues on Data Fusion, vol. 85, no. 1, pp. 6–23 (1997)
46. Hershey, J., Attias, H., Jojic, N., Krisjianson, T.: Audio visual graphical models for speech processing. In: IEEE International Conference on Speech, Acoustics, and Signal Processing, pp. 649–652. Montreal (2004)
47. Hershey, J., Movellan, J.: Audio-vision: using audio-visual synchrony to locate sounds. In: Advances in Neural Information Processing Systems, pp. 813–819. MIT Press, USA (2000)
48. Hinton, G.E., Osindero, S., Teh, Y.: A fast learning algorithm for deep belief nets. Neural Comput. **18**(7), 1527–1554 (2006)
49. Holzapfel, H., Nickel, K., Stiefelhagen, R.: Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3d pointing gestures. In: ACM International Conference on Multimodal Interfaces, pp. 175–182. State College, PA (2004)
50. Hossain, M.A., Atrey, P.K., El Saddik, A.: Smart mirror for ambient home environment. In: The 3rd IET International Conference on Intelligent Environments, pp. 589–596. Ulm (2007)
51. Hossain, M.A., Atrey, P.K., El Saddik, A.: Modeling and assessing quality of information in multi-sensor multimedia monitoring systems. ACM Trans. Multimed. Comput. Commun. Appl. **7**(1) (2011)
52. Hsu, W., Kennedy, L., Huang, C.W., Chang, S.F., Lin, C.Y.: News video story segmentation using fusion of multi-level multi-modal features in TRECVID 2003. In: International Conference on Acoustics Speech and Signal Processing. Montreal, QC (2004)
53. Hsu, W.H.M., Chang, S.F.: Generative, discriminative, and ensemble learning on multi-modal perceputal fusion toward news stroy segmentation. In: IEEE International Conference on Multimedia and Expos, pp. 1091–1094. Taipei (2004)
54. Hu, H., Gan, J.Q.: Sensors and data fusion algorithms in mobile robotics. Technical report, CSM-422, Department of Computer Science, University of Essex, UK (2005)
55. Hua, X.S., Zhang, H.J.: An attention-based decision fusion scheme for multimedia information retrieval. In: The 5th Pacific-Rim Conference on Multimedia. Tokyo, Japan (2004)
56. Isler, V., Bajcsy, R.: The sensor selection problem for bounded uncertainty sensing models. In: International Symposium on Information Processing in Sensor Networks, pp. 151–158. Los Angeles (2005)
57. Iyengar, G., Nock, H.J., Neti, C.: Audio-visual synchrony for detection of monologue in video archives. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. Hong Kong (2003)
58. Iyengar, G., Nock, H.J., Neti, C.: Discriminative model fusion for semantic concept detection and annotation in video. In: ACM International Conference on Multimedia, pp. 255–258. Berkeley (2003)
59. Jaffre, G., Pinquier, J.: Audio/video fusion: a preprocessing step for multimodal person identification. In: International Workshop on MultiModal User Authentification. Toulouse, France (2006)
60. Jaimes, A., Sebe, N.: Multimodal human computer interaction: a survey. In: IEEE International Workshop on Human Computer Interaction. Beijing (2005)
61. Jain, A., Nandakumar, K., Ross, A.: Score normalization in multimodal biometric systems. Pattern Recognit. **38**(12), 2270–2285 (2005)
62. Jasinschi, R.S., Dimitrova, N., McGee, T., Agnihotri, L., Zimmerman, J., Li, D., Louie, J.: A probabilistic layered framework for integrating multimedia content and context information. In: International Conference on Acoustics, Speech and Signal Processing, vol. II, pp. 2057–2060. Orlando (2002)
63. Jeon, J., Manmatha, R.: Using maximum entropy for automatic image annotation. In: International Conference on Image and Video Retrieval, vol. 3115, pp. 24–32. Dublin (2004)
64. Jiang, S., Kumar, R., Garcia, H.E.: Optimal sensor selection for discrete event systems with partial observation. IEEE Trans. Automat. Contr. **48**, 369–381 (2003)
65. Julier, S.J., Uhlmann, J.K.: New extension of the Kalman filter to nonlinear systems. In: Signal Processing, Sensor Fusion, and Target Recognition VI, vol. 3068 SPIE, pp. 182–193. San Diego (1997)
66. Kalman, R.E.: A new approach to linear filtering and prediction problems. Trans. ASME J. Basic Eng. **82**(series D), 35–45 (1960)
67. Kankanhalli, M.S., Wang, J., Jain, R.: Experiential sampling in multimedia systems. IEEE Trans. Multimed. **8**(5), 937–946 (2006)
68. Kankanhalli, M.S., Wang, J., Jain, R.: Experiential sampling on multiple data streams. IEEE Trans. Multimed. **8**(5), 947–955 (2006)
69. Kittler, J., Hatef, M., Duin, R.P., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **20**(3), 226–239 (1998)
70. Lam, K.Y., Cheng, R., Liang, B.Y., Chau, J.: Sensor node selection for execution of continuous probabilistic queries in wireless sensor networks. In: ACM International Workshop on Video Surveillance and Sensor Networks, pp. 63–71. NY, USA (2004)
71. León, T., Zuccarello, P., Ayala, G., de Ves, E., Domingo, J.: Applying logistic regression to relevance feedback in image retrieval systems. Pattern Recognit. **40**(10), 2621–2632 (2007)
72. Li, D., Dimitrova, N., Li, M., Sethi, I.K.: Multimedia content processing through cross-modal association. In: ACM International Conference on Multimedia (2003)
73. Li, F.F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. **2**, pp. 524–531. Washington (2005)
74. Li, M., Li, D., Dimitrove, N., Sethi, I.K.: Audio-visual talking face detection. In: International Conference on Multimedia and Expo, pp. 473–476. Baltimore, MD (2003)
75. Liu, X., Zhang, L., Li, M., Zhang, H., Wang, D.: Boosting image classification with lda-based feature combination for digital photograph management. Pattern Recognit. **38**(6), 887–901 (2005)
76. Liu, Y., Zhang, D., Lu, G., Tan, A.H.: Integrating semantic templates with decision tree for image semantic learning. In: The 13th International Multimedia Modeling Conference, pp. 185–195. Singapore (2007)
77. Loh, A., Guan, F., Ge, S.S.: Motion estimation using audio and video fusion. In: International Conference on Control, Automation, Robotics and Vision, vol. 3, pp. 1569–1574 (2004)
78. Lucey, S., Sridharan, S., Chandran, V.: Improved speech recognition using adaptive audio-visual fusion via a stochastic secondary classifier. In: International Symposium on Intelligent Multimedia, Video and Speech Processing, pp. 551–554. Hong Kong (2001)
79. Luo, R.C., Yih, C.C., Su, K.L.: Multisensor fusion and integration: Approaches, applications, and future research directions. IEEE Sens. J. **2**(2), 107–119 (2002)
80. Magalhães, J., Rüger, S.: Information-theoretic semantic multimedia indexing. In: International Conference on Image and

Video Retrieval, pp. 619–626. Amsterdam, The Netherlands (2007)
81. Makkook, M.A.: A multimodal sensor fusion architecture for audio-visual speech recognition. MS Thesis, University of Waterloo, Canada (2007)
82. Matas, J., Hamouz, M., Jonsson, K., Kittler, J., Li, Y., Kotropoulos, C., Tefas, A., Pitas, I., Tan, T., Yan, H., Smeraldi, F., Capdevielle, N., Gerstner, W., Abdeljaoued, Y., Bigun, J., Ben-Yacoub, S., Mayoraz, E.: Comparison of face verification results on the XM2VTS database. p. 4858. Los Alamitos, CA, USA (2000)
83. McDonald, K., Smeaton, A.F.: A comparison of score, rank and probability-based fusion methods for video shot retrieval. In: International Conference on Image and Video Retrieval, pp. 61–70. Singapore (2005)
84. Mena, J.B., Malpica, J.: Color image segmentation using the dempster–shafer theory of evidence for the fusion of texture. In: International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XXXIV, Part 3/W8, pp. 139–144. Munich, Germany (2003)
85. Meyer, G.F., Mulligan, J.B., Wuerger, S.M.: Continuous audio-visual digit recognition using N-best decision fusion. J. Inf. Fusion **5**, 91–101 (2004)
86. Nefian, A.V., Liang, L., Pi, X., Liu, X., Murphye, K.: Dynamic bayesian networks for audio-visual speech recognition. EURASIP J. Appl. Signal Process. **11**, 1–15 (2002)
87. Neti, C., Maison, B., Senior, A., Iyengar, G., Cuetos, P., Basu, S., Verma, A.: Joint processing of audio and visual information for multimedia indexing and human-computer interaction. In: International Conference RIAO. Paris, France (2000)
88. Ni, J., , Ma, X., Xu, L., Wang, J.: An image recognition method based on multiple bp neural networks fusion. In: IEEE International Conference on Information Acquisition (2004)
89. Nickel, K., Gehrig, T., Stiefelhagen, R., McDonough, J.: A joint particle filter for audio-visual speaker tracking. In: The 7th International Conference on Multimodal Interfaces, pp. 61–68. Torento, Italy (2005)
90. Nock, H.J., Iyengar, G., Neti, C.: Assessing face and speech consistency for monologue detection in video. In: ACM International Conference on Multimedia. French Riviera, France (2002)
91. Nock, H.J., Iyengar, G., Neti, C.: Speaker localisation using audio-visual synchrony: an empirical study. In: International Conference on Image and Video Retrieval. Urbana, USA (2003)
92. Noulas, A.K., Krose, B.J.A.: Em detection of common origin of multi-modal cues. In: International Conference on Multimodal Interfaces, pp. 201–208. Banff (2006)
93. Ortega-Garcia, J., Fierrez-Aguilar, J., Simon, D., Gonzalez, J., Faundez-Zanuy, M., Espinosa, V., Satue, A., Hernaez, I., Igarza, J.J., Vivaracho, C., Escudero, D., Moro, Q.I.: Biometric on the internet MCYT baseline corpus: a bimodal biometric database. IEE Proc. Vis. Image Signal Process. **150**(6), 395–401 (2003)
94. Oshman, Y.: Optimal sensor selection strategy for discrete-time state estimators. IEEE Trans. Aerosp. Electron. Syst. **30**, 307–314 (1994)
95. Oviatt, S.: Ten myths of multimodal interaction. Commun. ACM **42**(11), 74–81 (1999)
96. Oviatt, S.: Taming speech recognition errors within a multimodal interface. Commun. ACM **43**(9), 45–51 (2000)
97. Oviatt, S.L.: Multimodal interfaces. In: Jacko, J., Sears, A. (eds.) The Human–Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications. Lawrence Erlbaum Assoc., NJ (2003)
98. Pahalawatta, P., Pappas, T.N., Katsaggelos, A.K.: Optimal sensor selection for video-based target tracking in a wireless sensor network. In: IEEE International Conference on Image Processing, pp. V:3073–3076. Singapore (2004)
99. Perez, D.G., Lathoud, G., McCowan, I., Odobez, J.M., Moore, D.: Audio-visual speaker tracking with importance particle filter. In: IEEE International Conference on Image Processing (2003)
100. Pfleger, N.: Context based multimodal fusion. In: ACM International Conference on Multimodal Interfaces, pp. 265–272. State College (2004)
101. Pfleger, N.: Fade - an integrated approach to multimodal fusion and discourse processing. In: Dotoral Spotlight at ICMI 2005. Trento, Italy (2005)
102. Pitsikalis, V., Katsamanis, A., Papandreou, G., Maragos, P.: Adaptive multimodal fusion by uncertainty compensation. In: Ninth International Conference on Spoken Language Processing. Pittsburgh (2006)
103. Poh, N., Bengio, S.: How do correlation and variance of base-experts affect fusion in biometric authentication tasks? IEEE Trans. Signal Process. **53**, 4384–4396 (2005)
104. Poh, N., Bengio, S.: Database, protocols and tools for evaluating score-level fusion algorithms in biometric authentication. Pattern Recognit. **39**(2), 223–233 (2006) (Part Special Issue: Complexity Reduction)
105. Potamianos, G., Luettin, J., Neti, C.: Hierarchical discriminant features for audio-visual LVSCR. In: IEEE International Conference on Acoustic Speech and Signal Processing, pp. 165–168. Salt Lake City (2001)
106. Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A.: Recent advances in the automatic recognition of audiovisual speech. Proc. IEEE **91**(9), 1306–1326 (2003)
107. Potamitis, I., Chen, H., Tremoulis, G.: Tracking of multiple moving speakers with multiple microphone arrays. IEEE Trans. Speech Audio Process. **12**(5), 520–529 (2004)
108. Radova, V., Psutka, J.: An approach to speaker identification using multiple classifiers. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, **2**, 1135–1138. Munich, Germany (1997)
109. Rashidi, A., Ghassemian, H.: Extended dempster–shafer theory for multi-system/sensor decision fusion. In: Commission IV Joint Workshop on Challenges in Geospatial Analysis, Integration and Visualization II, pp. 31–37. Germany (2003)
110. Reddy, B.S.: Evidential reasoning for multimodal fusion in human computer interaction (2007). MS Thesis, University of Waterloo, Canada
111. Ribeiro, M.I.: Kalman and extended Kalman filters: concept, derivation and properties. Technical report., Institute for Systems and Robotics, Lisboa (2004)
112. Roweis, S., Ghahramani, Z.: A unifying review of linear gaussian models. Neural Comput. **11**(2), 305–345 (1999)
113. Sanderson, C., Paliwal, K.K.: Identity verification using speech and face information. Digit. Signal Process. **14**(5), 449–480 (2004)
114. Satoh, S., Nakamura, Y., Kanade, T.: Name-It: Naming and detecting faces in news video. IEEE Multimed. **6**(1), 22–35 (1999)
115. Siegel, M., Wu, H.: Confidence fusion. In: IEEE International Workshop on Robot Sensing, pp. 96–99 (2004)
116. Singh, R., Vatsa, M., Noore, A., Singh, S.K.: Dempster–shafer theory based finger print classifier fusion with update rule to minimize training time. IEICE Electron. Express **3**(20), 429–435 (2006)
117. Slaney, M., Covell, M.: Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. In: Neural Information Processing Society, vol. 13 (2000)
118. Smeaton, A.F., Over, P., Kraaij, W.: High-level feature detection from video in TRECVid: a 5-year retrospective of achievements. In: Divakaran, A. (ed.) Multimedia Content Analysis, Theory and Applications, pp. 151–174. Springer, Berlin (2009)

119. Snoek, C.G.M., Worring, M.: A review on multimodal video indexing. In: IEEE International Conference on Multimedia and Expo, pp. 21–24. Lusanne, Switzerland (2002)
120. Snoek, C.G.M., Worring, M.: Multimodal video indexing: A review of the state-of-the-art. Multimed. Tools Appl. **25**(1), 5–35 (2005)
121. Snoek, C.G.M., Worring, M., Smeulders, A.W.M.: Early versus late fusion in semantic video analysis. In: ACM International Conference on Multimedia, pp. 399–402. Singapore (2005)
122. Sridharan, H., Sundaram, H., Rikakis, T.: Computational models for experiences in the arts and multimedia. In: The ACM Workshop on Experiential Telepresence. Berkeley, CA (2003)
123. Stauffer, C.: Automated audio-visual activity analysis. Tech. rep., MIT-CSAIL-TR-2005-057, Massachusetts Institute of Technology, Cambridge, MA (2005)
124. Strobel, N., Spors, S., Rabenstein, R.: Joint audio–video object localization and tracking. IEEE Signal Process. Mag. **18**(1), 22–31 (2001)
125. Talantzis, F., Pnevmatikakis, A., Polymenakos, L.C.: Real time audio-visual person tracking. In: IEEE 8th Workshop on Multimedia Signal Processing, pp. 243–247. IEEE Computer Society, Victoria, BC (2006)
126. Tatbul, N., Buller, M., Hoyt, R., Mullen, S., Zdonik, S.: Confidence-based data management for personal area sensor networks. In: The Workshop on Data Management for Sensor Networks (2004)
127. Tavakoli, A., Zhang, J., Son, S.H.: Group-based event detection in undersea sensor networks. In: Second International Workshop on Networked Sensing Systems. San Diego, CA (2005)
128. Teissier, P., Guerin-Dugue, A., Schwartz, J.L.: Models for audiovisual fusion in a noisy-vowel recognition task. J. VLSI Signal Process. **20**, 25–44 (1998)
129. Teriyan, V.Y., Puuronen, S.: Multilevel context representation using semantic metanetwork. In: International and Interdisciplinary Conference on Modeling and Using Context, pp. 21–32. Rio de Janeiro, Brazil (1997)
130. Tesic, J., Natsev, A., Lexing, X., Smith, J.R.: Data modeling strategies for imbalanced learning in visual search. In: IEEE International Conference on Multimedia and Expo, pp. 1990–1993. Beijing (2007)
131. Town, C.: Multi-sensory and multi-modal fusion for sentient computing. Int. J. Comput. Vis. **71**, 235–253 (2007)
132. Vermaak, J., Gangnet, M., Blake, A., Perez, P.: Sequential monte carlo fusion of sound and vision for speaker tracking. In: The 8th IEEE International Conference on Computer Vision, vol. 1, pp. 741–746. Paris, France (2001)
133. Voorhees, E.M., Gupta, N.K., Johnson-Laird, B.: Learning collection fusion strategies. In: ACM International Conference on Research and Development in Information Retrieval, pp. 172–179. Seattle, WA (1995)
134. Wall, M.E., Rechtsteiner, A., Rocha, L.M.: Singular Value Decomposition and Principal Component Analysis, Chap. 5, pp. 91–109. Kluwel, Norwell, MA (2003)
135. Wang, J., Kankanhalli, M.S.: Experience-based sampling technique for multimedia analysis. In: ACM International Conference on Multimedia, pp. 319–322. Berkeley, CA (2003)
136. Wang, J., Kankanhalli, M.S., Yan, W.Q., Jain, R.: Experiential sampling for video surveillance. In: ACM Workshop on Video Surveillance. Berkeley (2003)
137. Wang, S., Dash, M., Chia, L.T., Xu, M.: Efficient sampling of training set in large and noisy multimedia data. ACM Trans. Multimed. Comput. Commun. Appl. **3**(3), 14 (2007)
138. Wang, Y., Liu, Z., Huang, J.C.: Multimedia content analysis: using both audio and visual clues. In: IEEE Signal Processing Magazine, pp. 12–36 (2000)
139. Westerveld, T.: Image retrieval: content versus context. In: RIAO Content-Based Multimedia Information Access. Paris, France (2000)
140. Wu, H.: Sensor data fusion for context-aware computing using dempster–shafer theory. Ph.D. thesis, The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA (2003)
141. Wu, K., Lin, C.K., Chang, E., Smith, J.R.: Multimodal information fusion for video concept detection. In: IEEE International Conference on Image Processing, pp. 2391–2394. Singapore (2004)
142. Wu, Y., Chang, E., Tsengh, B.L.: Multimodal metadata fusion using causal strength. In: ACM International Conference on Multimedia, pp. 872–881. Singapore (2005)
143. Wu, Y., Chang, E.Y., Chang, K.C.C., Smith, J.R.: Optimal multimodal fusion for multimedia data analysis. In: ACM International Conference on Multimedia, pp. 572–579. New York City, NY (2004)
144. Wu, Z., Cai, L., Meng, H.: Multi-level fusion of audio and visual features for speaker identification. In: International Conference on Advances in Biometrics, pp. 493–499 (2006)
145. Xie, L., Kennedy, L., Chang, S.F., Divakaran, A., Sun, H., Lin, C.Y.: Layered dynamic mixture model for pattern discovery in asynchronous multi-modal streams. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 1053–1056. Philadelphia, USA (2005)
146. Xiong, N., Svensson, P.: Multi-sensor management for information fusion: issues and approaches. Inf. Fusion **3**, 163–186(24) (2002)
147. Xu, C., Wang, J., Lu, H., Zhang, Y.: A novel framework for semantic annotation and personalized retrieval of sports video. IEEE Trans. Multimed. **10**(3), 421–436 (2008)
148. Xu, C., Zhang, Y.F., Zhu, G., Rui, Y., Lu, H., Huang, Q.: Using webcast text for semantic event detection in broadcast sports video. IEEE Trans. Multimed. **10**(7), 1342–1355 (2008)
149. Xu, H., Chua, T.S.: Fusion of AV features and external information sources for event detection in team sports video. ACM Trans. Multimed. Comput. Commun. Appl. **2**(1), 44–67 (2006)
150. Yan, R.: Probabilistic models for combining diverse knowledge sources in multimedia retrieval. Ph.D. thesis. Carnegie Mellon University (2006)
151. Yan, R., Yang, J., Hauptmann, A.: Learning query-class dependent weights in automatic video retrieval. In: ACM International Conference on Multimedia, pp. 548–555. New York, USA (2004)
152. Yang, M.T., Wang, S.C., Lin, Y.Y.: A multimodal fusion system for people detection and tracking. International Journal of Imaging Systems and Technology **15**, 131–142 (2005)
153. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A. Face recognition: a literature survey. ACM Comput. Surv. **35**(4), 399–458 (2003)
154. Zhou, Q., Aggarwal, J.: Object tracking in an outdoor environment using fusion of features and cameras. Image Vis. Comput. **24**(11), 1244–1255 (2006)
155. Zhou, Z.H.: Learning with unlabeled data and its application to image retrieval. In: The 9th Pacific Rim International Conference on Artificial Intelligence, pp. 5–10. Guilin (2006)
156. Zhu, Q., Yeh, M.C., Cheng, K.T.: Multimodal fusion using learned text concepts for image categorization. In: ACM International Conference on Multimedia, pp. 211–220. Santa Barbara (2006)
157. Zotkin, D.N., Duraiswami, R., Davis, L.S.: Joint audio-visual tracking using particle filters. EURASIP J. Appl. Signal Process. (11), 1154–1164 (2002)
158. Zou, X., Bhanu, B.: Tracking humans using multimodal fusion. In: IEEE Conference on Computer Vision and Pattern Recognition, p. 4. Washington (2005)