P. Maragos, P. Gros, A. Katsamanis and G. Papandreou,
"Cross-Modal Integration for Performance Improving in Multimedia: A Review",
Chapter in Multimodal Processing and Interaction: Audio, Video, Text,
edited by P. Maragos, A. Potamianos and P. Gros, Springer-Verlag, 2008.

1

Cross-Modal Integration for Performance Improving in Multimedia: A Review

Petros Maragos[1], Patrick Gros[2], Athanassios Katsamanis[1], and George Papandreou[1]

[1] National Technical University of Athens, Greece
[2] Institut National de Recherche en Informatique et Automatique, France

Our surrounding world is abundant with multimodal stimuli which emit multisensory information in the form of analog signals. Humans perceive the natural world in a multimodal way: vision, hearing, touch. Nowadays, propelled by our digital technology, we are also witnessing a rapid explosion of digital multimedia data. Humans understand the multimodal world in a seemingly effortless manner, although there are vast information processing resources dedicated to the corresponding tasks by the brain. Computer techniques, despite recent advances, still significantly lag humans in understanding multimedia and performing high-level cognitive tasks. Some of these limitations are inborn, i.e., stem from the complexity of the data and their multimodality. Other shortcomings, though, are due to the inadequacy of most approaches used in multimedia analysis, which are essentially monomodal. Namely, they rely mainly on information from a single modality and on tools effective for this modality while they underutilize the information in other modalities and their cross-interaction. To some extent, this happens because most researchers and groups are still monomedia specialists. Another reason is that the problem of fusing the modalities has not still reached maturity, both from a mathematical modeling and a computational viewpoint. Consequently, a major scientific and technological challenge is to develop truly multimodal approaches that integrate several modalities toward improving the goals of multimedia understanding. In this chapter we review research on the theory and applications of several multimedia analysis approaches that improve robustness and performance through cross-modal integration.

1.1 Motivations and Problems

Digital technology provides us with multimedia data whose size and complexity keeps rapidly expanding. To analyze and understand them we must face major challenges which include the following:

Data are Voluminous: Nowadays we are witnessing a rapid explosion of digital multimedia data. They are produced by a variety of sources including: video cameras, TV, digital photography (personal and professional albums, photo agencies), digital audio and other digital entertainment devices, digital audiovisual libraries, multimodal Web. As a numeric example, 24 hr of TV produces 430 Gb (raw, uncompressed) data, 2.160.000 still (frame) images.

Data are Dynamic: Dynamic websites, TV and other broadcast news quickly get obsolete.

Different Temporal Rates are of importance in the various media. For example, 25-30 image-frames/sec in video, 44.000 sound samples/sec in audio, 100 feature-frames/sec in speech, 4 syllables/sec in language processing.

Cross-Media asynchrony, since image and audio scene boundaries may be different. Examples include possible asynchrony between the voice heard and the face seen, or between a sports visual event (e.g., a goal in soccer) and the speaker's comment that comes later.

Monomedia specialization: Most researchers and groups are specialists in a single modality, e.g., speech processing and recognition, or image/video processing and computer vision, or natural language processing.

The rapid explosion of multimedia data creates an increasing difficulty in finding relevant information, which has spurred enormous efforts to develop tools for automatic detection, recognition, and semantic analysis of multimedia content. The overall goal is multimedia understanding, which requires to use content in a nontrivial way. For example, understanding goes beyond just displaying images or playing a music CD, for which we do not need to analyze the content of the data. In contrast, examples that require understanding include multimedia archiving, re-purposing, making websites from TV streams. This multimedia explosion also poses the need to develop efficient solutions for problems in several ambitious technology areas. Two such grand challenges[3] are: (i) Natural access and high-level interaction with multimedia databases, and (ii) Detecting, recognizing and interpreting objects, events, and human behavior in multimedia videos by processing combined audio-video-text data.

Thus, as mentioned in this book's Introduction, one of the grand goals in multimedia understanding is cross-media integration for improving performance in the various scientific and technological problems that arise in systems dealing with multiple modalities. And this is exactly the central topic of this review chapter. Among the features of this chapter, we include brief reviews of ideas and results from cross-modal integration in human perception, since the multimodal human brain is a great source of inspiration. Further, we survey several types of probabilistic approaches and models for multimodal fusion. Examples of modalities to integrate include combinations of:

- vision and speech/audio

[3] These challenges were also identified at `http://www.muscle-noe.org/`

- vision (or speech) and tactile
- image/video (or speech/audio) and text
- multiple-cue versions of vision and/or speech
- other semantic information or metadata.

Many previous research efforts in (human or machine) cross-modal integration deal with combining multiple cues, i.e., multiple streams of information from the same modality. A frequent example is vision, where multiple cues are often combined to increase the robustness in estimating properties of the visual world scene; e.g., stereo disparity is combined with texture to estimate depth. In general, if we wish to refine the definition of multimodality, we shall call **multicue** the *intramodal* integration of several cues within the same modality and **multimodal** the *intermodal* integration of several modalities. For example, to estimate the depth of object surfaces by combining stereo and texture is a multicue approach, whereas combining vision and haptics is a multimodal approach. However, for expressional simplicity, we may often use only the term 'multimodal' to refer to both intermodal and intramodal approaches.

The combinations of modalities (or cues) can be either of the cross-interaction type or of the cross-integration type. Interaction implies an information reaction-diffusion among modalities with feedback control of one modality by others. Integration involves exploiting heterogeneous information cumulatively from various modalities in a data feature fusion toward improved performance. A simpler way to see this differentiation is to consider strong- versus weak-coupling of modalities (discussed later in Section 1.3). Some broad areas of research problems in multimedia where integration of (strongly-coupled or weakly-coupled) modalities occurs include the following:

- **Features:** The extraction of critical features in each modality, e.g., audio, vision, text, is in a well-advanced state and is served by the fields of signal processing and pattern recognition. See Section 1.4 for a brief survey. However, when combining several modalities, it is quite challenging to integrate monomodal[4] features in a way that is robust (since individual stream reliabilities may vary dynamically), efficient in terms of size and synchrony, and optimum in terms of overall performance. Thus, some ongoing research challenges in this classic problem of multimedia include: (i) Selection, robust extraction, and dimensionality reduction of each modality's features, given the presence of other modalities and their corresponding features. (ii) Optimal fusion of the separate feature streams (from different modalities or cues). A typical example is the area of audio-visual speech recognition, where the audio feature extraction has advanced, but there is still ongoing research for robust extraction of low-dimensional visual speech features and optimal fusion of the audio and visual features.

[4] In this chapter, the term monomodal is used as equivalent to unimodal.

- **Models:** Most aspects of multimedia understanding involve problems in pattern recognition and machine learning. One can select appropriate methodologies and algorithms from the vast arena of these fields, including both static and dynamic classification models. However, in multimodal processing and integration, the big challenge is how to adapt or extend these models so that they can work with and decide optimally for multimodal data. For instance, an important issue is whether to fuse the data at an early, intermediate, or late stage of the integration procedure. Another challenge is to deal with the time-dependent nature of these data when the modalities are not synchronous. These issues are discussed in Section 1.5.
- **Applications:** The application areas of multimedia are numerous and keep growing. Examples that involve cross-modal integration include the following: (See also Section 1.7 for a brief survey of some applications.)
  - *Audiovisual Speech:* The two problems of automatically recognizing speech and inverting speech, i.e., recovering the geometry of the vocal tract, are ill-posed. Integrating the auditory information with visual information (e.g., video features around the speaker's mouth area) imposes additional constraints which may help regularizing the solution of these problems.
  - *Cross-Media Interaction Scenarios in Human Computer Interfaces (HCIs):* Human-computer interaction has started becoming a reality due to recent advances in speech recognition, natural language processing, object detection-tracking using vision and tactile sensors. However, building a natural and efficient HCI that combines all the required different modalities (e.g., speech, vision, graphics, text, tactile) toward improving the overall performance becomes a significant technical challenge in this case where the modalities can interact strongly. A review of this area is given in the book's Chapter **??**.
  - *Multimodal Saliency:* Audiovisual Attention Modeling and Salient Event Detection is a significant research problem with applications in audiovisual event detection, bimodal emotion recognition, dialogue detection, and video summarization. A significant effort in this area is spent on multimodal feature extraction and fusion for attention modeling. (See Chapter **??**.)
  - *Video Analysis and Integration of Asynchronous Time-Evolving Modalities:* Video processing is usually done separately on sound and on images. However, the solution of many video analysis tasks can be improved and become more robust by integrating these two modalities and possibly text. Major difficulties exist, however, because the various media are not temporally coherent and provide different kinds of data. Several chapters in this book deal with these problems.
  - *Combining Text and Vision or Text and Audio for Semantic Labeling:* The challenging research goal here is to use structural and textual information for semantic interpretation of image or audio data. Such

technologies will empower a full semantic analysis and classification of data for which currently almost everything beyond text is ignored.

The areas of human or machine cross-modal integration are both huge, and hence our coverage in this chapter will *not* be exhaustive. Instead, we sample key ideas and survey indicative applications. The rest of this chapter is organized as follows. In Section 1.2 we briefly summarize how various branches of psychology view perception, how certain aspects of perceptual inference can be modeled via Bayesian estimation and decision theory, and then we present examples of multicue or multimodal perception from psychophysics. In Section 1.3 we classify sensor data fusion schemes using a Bayesian formulation. The following four sections review the main problem areas in multimedia analysis and integration: feature extraction from the three main modalities (speech-audio, image-video, and text) in Section 1.4; stochastic models for cross-modal integration in Section 1.5; integrated multimedia content analysis beyond descriptors in Section 1.6; and a few sample applications areas in Section 1.7. Finally, we conclude in Section 1.8 by outlining promising future directions.

1.2 Multimodality in Human Perception

Humans need to extract multi-level information about the structures and their spatio-temporal or cognitive relationships in their world environment. This information processing could either be innate (inborn) and possibly learned via evolutionary processes or stimulated by sensory data. This chapter mainly focuses on the latter. The polarity between innate vs data-driven inference is conceptually similar to (or inspired by) Plato's rationalism versus Aristotle's empiricism.

Three stages in sensory information processing are sensation, perception and cognition. **Sensation** is signal formation caused by the sense organs (i.e., the sensors) when excited by the external world stimuli. **Perception** is the collection of processes by which we filter, select, organize, recognize, and understand the sensations. There is an overlap between sensation and perception, but as broadly stated in [38], "sensations are usually viewed as simple, basic experiences caused by simple stimuli, whereas perceptions are usually considered as more complicated experiences elicited by complex, often meaningful, stimuli". Even more complicated is **cognition** which refers to information analyzing mental processes such as comprehension, learning, memory, decision-making, planning. A causal hierarchy may be the following:

$$\text{Sensation} \longrightarrow \text{Perception} \longrightarrow \text{Cognition}$$

Since the dividing line is usually hard to draw between sensations and perceptions as well as between perception and cognition, henceforth, we shall loosely refer to *perception as the sensory-based inference about the world state*, i.e.,

the process through which the mapping from world stimuli to sensory signals is inverted. Herein, inference is meant broadly as the collection of the main tasks of sensory information processing, e.g., spatio-temporal detection of objects and events, estimation of their properties, localization, recognition, organization.

Human perception as a means of daily exploration and survival in nature has been of vital importance since the dawn of humanity. As a physical process or result of sensor operation, it has attracted the interest of great scientists in the physical sciences (acoustics, optics, neurobiology). As a main ingredient of human awareness and consciousness, its understanding has also occupied the minds of great philosophers, artists and psychologists. Approaches to study perception range from physiology and neurobiology through cognition-related psychology disciplines to philosophy disciplines centered around the mind-body problem. A practical blend of the first two viewpoints is presented by *psychophysics*, a subdiscipline of psychology, which explores the relationships between the external world's physical stimuli and their induced percepts in the human mind.

In the rest of this section we summarize how various branches of psychology view perception, how certain aspects of perceptual inference can be modeled via Bayesian estimation and decision theory, and then we present examples of multicue or multimodal perception from psychophysics. Obviously, since this is a huge area, here we only summarize some indicative cases that have proven useful in monomodal or multimodal information processing.

1.2.1 Psychology Approaches to Human Perception

For the aspects of sensory-based human perception that we will need in this review chapter on multimodal integration, most important are the disciplines of gestalt psychology and cognitive psychology. Before we summarize their main ideas, we outline a few of their origins from philosophy.

Much like the mind-body debate, ideas and approaches in psychology evolved from the poles of philosophy and physiology. The former relies primarily on reasoning and introspection, whereas the latter on empirical methods and observations. As in other sciences, the evolution of ideas in philosophy [96] often followed a *dialectic* path, where a new theory was proposed (a *thesis*), soon countered by an opposite theory (an *antithesis*), until a *synthesis* of the best ideas was formed. This synthesis formed a new thesis, to be followed by a new antithesis, and so on dialectically. A classic pair of thesis and antithesis is Plato's *rationalism* versus Aristotle's *empiricism*. In the former we are supposed to acquire most knowledge mainly via theoretical analysis (understanding and reasoning) independently of the senses, in the latter mainly via empirical evidence (experience and observations, especially sensory perception) independently of innate ideas. A similar contrasting controversy continued in modern philosophy between the rationalist Descartes, whose assertion "I think, therefore I am" cannot be doubted and views the

mind as more certain than matter, versus the empiricist Locke, who emphasized experience and learning and believed that everything knowable (with the possible exception of logic and mathematics) is derived from perception. Kant synthesized both their views. Such a synthesis is used in most modern theories of knowledge, where elements from both rationalism and empiricism are encountered.

Along the route of empiricism, *behaviorism* in psychology developed as a discipline that focuses on observable behaviors as responses to the environment, without any consideration to internal processes or mind theories. According to behaviorism, all what humans do, both externally (actions) and internally (thoughts), are behaviors.

An avid rival to behaviorism, **Gestalt psychology** is a mind-brain theory for which the most important process is the formation of perceptual groups of observations that correspond to conceptual equivalence classes. An abstraction of the justification for preferring this grouping is the Latin adage *multum non multa*, which distinguishes two meanings of 'much': The former emphasizes how a deeper understanding can grow from analyzing and grouping of fewer items, whereas the latter is based on quantitative detailed analysis of many data. Thus, the gestalt theory is a global, holistic approach (as opposed to the local, atomistic observations of behaviorism). It is concerned with *molar* behavior instead of molecular behavior, the former being a coarse-granule grouping of behavior in external settings, whereas the latter is the fine-granule behavior taking place internally inside an organism initiated by environmental stimuli. The gestalt thesis that the whole is greater than the sum of its parts is particularly relevant for multimodal processing. It implies that grouping in the sense of fusing modalities creates a unifying percept that subsumes their simple concatenation.

Founded by Wertheimer, Köhler and Koffka [58, 59] during 20th century's first half, gestalt psychology distinguishes between the geographical environment versus the behavioral environment and emphasizes that perception occurs in the latter. However, the behavioral environment B by itself is not sufficient to account for all processes and needs to be complemented by the physiological processes P active during perception. B and P are psychophysically isomorphic. Wertheimer's principle of *psychophysical isomorphism* is to think of physiological processes not as molecular but as molar phenomena. Köhler [59] refined this principle and proposed the following in the cases of spatial and temporal order: "(i) Experienced order in space is always structurally identical with a functional order in the distribution of underlying brain processes. (ii) Experienced order in time is always structurally identical with a functional order in the sequence of correlated brain processes."

The main ideas in gestalt psychology have been inspired by or geared toward problems in visual perception. The perceptual grouping forms objects by starting from local data or features who satisfy or share several grouping principles and recursively builds larger visual objects, the Gestalts. The most important of these principles is the law of Prägnanz, according to which we

perceive a collection of visual parts in the simplest way that can organize the partial elements into a stable and coherent form. Other gestalt principles include proximity, figure-ground, closure, continuity, similarity, and symmetry. Additional characteristics of the gestalt theory are that, it focuses on a parallel and continuous processing and favors self-organization.

An outgrowth of gestalt psychology and Piaget's stage theory for child cognitive development is the field of **Cognitive psychology**, which (according to Neisser who introduced the term in 1967) is "the study of how people learn, structure, store and use knowledge." A comprehensive introduction can be found in [110]. This school of psychology is concerned with how humans process information for general tasks such as perception, learning, memory, language, problem-solving. Unlike behaviorism, it accepts innate mental states, but it also uses scientific methods of experimentation and observation without resorting to introspection. Due to its emphasis on the mental processes associating stimuli and responses, it uses computational concepts, like input and output of mental processes, algorithms and knowledge representation. As such, it is closer to artificial intelligence, and the two fields have benefited from cross-fertilization. Actually, cognitive psychology has contributed to artificial intelligence the very useful concept and tool of semantic networks. For example, WordNet [33] is a semantic network used in natural language processing.

The most often used practical tools to test gestalt and cognitive theories of human perception stem from psychophysics and statistics. The psychophysical methods deal with determination of sensory thresholds, measurements of sensitivity, and signal detection theory. From statistics, the Bayesian framework has gained popularity and is briefly summarized next.

1.2.2 Bayesian Formulation of Perception

Bayesian statistics provides a general framework for modeling and solving problems in pattern recognition and machine learning [29, 12, 112] and in computer vision [39, 37, 67, 17]. Its success in vision has also propelled its use for modeling perception as Bayesian inference [57, 124]. Elements of the Bayesian framework for perception can be found in Helmholtz's belief that retinal images alone do not carry sufficient information and need to be supplemented with prior knowledge; hence, he viewed perception as unconscious inference [41]. Although the Bayesian approach to perception has been mainly developed for vision, we shall use herein the Bayesian formalism to model multimodal sensory information processing, where 'multimodal' may mean 'multi-sensory'. A unifying Bayesian approach to computer vision, speech and signal processing and their associated pattern analysis and recognition tasks is also offered by the 'Pattern Theory' field [39, 75].

For intuition and simplicity, we will often restrict to the two main senses, vision and hearing, and use the term 'audiovisual' instead of 'multimodal sensory' stimuli/data. Let S be a configuration (of the properties) of an auditory

and/or visual scene of the external world (e.g., a vector of variables representing numeric or symbolic features-parameters) that represents the state of external audiovisual stimuli. Let D be the monomodal or multimodal data representing signals collected by auditory and/or visual sensors; at a higher level, D may also represent features extracted from the raw signals. If we view the sensory signal formation as a mapping $S \mapsto D$ from world state S to sensory data D, then perception is essentially the inverse problem of estimating the world audiovisual state from the sensory observations. If the variations of the audiovisual state are random in nature, or there is uncertainty in modeling the signal formation or there is observation noise, then we can use a probabilistic interpretation of the above problem. In this case, Bayes' formula offers us a convenient decomposition of the probabilities involved into prior (before observing the data) and posterior (after observing the data) terms:

$$P(S|D) = \frac{P(D|S)P(S)}{P(D)} \tag{1.1}$$

where $P(\cdot)$ denotes probability distributions (i.e., probability densities or probability masses according to the case). The **prior** distribution $P(S)$ expresses the *a priori* probability of how likely is the world state S before observing the data; it models prior knowledge about the random nature (e.g. regularities) of the scene structure and may include various a priori constraints. The conditional distribution $P(D|S)$ expresses the probability of observing D given the world state S; if it is viewed as function of S for fixed D, then it is called the **likelihood** function of S. It statistically models the overall causal generation process of signal data formation from the world state (audiovisual scene); thus, this probabilistic mapping $S \mapsto D$ is called a *generative model* in Bayesian networks. The likelihood embodies the reliability of the observed signal or feature data D which can vary due to possible model uncertainty and observation noise. The *marginal* distribution $P(D)$, usually called the **evidence**, expresses the probability of observing the data under all mutually exclusive state configurations; it can be computed by summing the product of the likelihood times the prior over all such S. Herein, we shall assume that the world state variables vary continuously and hence $P(D) = \int P(D|S)P(S)dS$. The distribution $P(D)$ encapsulates data regularities that arise from similarities among audiovisual scenes in nature. Finally, the **posterior** conditional distribution $P(S|D)$ expresses the *a posteriori* probability of the audiovisual scene S after observing the data D.

The posterior distribution is the main tool for Bayesian inference since it allows us to use the data as observations to update the estimate of S based on Bayes' formula. This updating, applied to perception, agrees with cognitive psychology's view that, as we move in the environment we sense the world and our sensations get mapped to percepts which are accompanied by degrees of belief; these percepts may change as we acquire new information. In addition to the posterior, to complete the inference process, we also need a *decision rule*. For example, one of the most well-known solutions for finding S is to

select the *Maximum-A-Posteriori (MAP)* estimate:

$$\hat{S}_{MAP} = \underset{S}{\operatorname{argmax}}\, P(D|S)P(S) \tag{1.2}$$

The marginal $P(D)$ is viewed as a normalization factor and hence is ignored in this maximization. The MAP estimate is influenced both from prior knowledge and from the data observed. Thus, assuming a uniform prior reduces the above to the equally well-known *Maximum Likelihood (ML)* estimate

$$\hat{S}_{ML} = \underset{S}{\operatorname{argmax}}\, P(D|S) \tag{1.3}$$

A unifying way to view these and other solutions is through Bayesian **decision** theory. First, we specify a *loss* (negative utility) function $L(S, A)$ that associates a cost L to the *decision* that assigns a solution A to the true scene state S. The *risk* is the expected loss over all possible scenes:

$$\text{Risk}(A) = \int L(S, A)P(S|D)dS \tag{1.4}$$

Then, we find an optimum Bayesian decision, i.e., solution $\hat{S}$, by minimizing this risk:

$$\hat{S} = \underset{A}{\operatorname{argmin}}\, \text{Risk}(A) \tag{1.5}$$

If we set $L(S, A) = c - \delta(S - A)$ where δ is the Dirac function, which means that we penalize equally (by a cost c) all wrong decisions, then $\text{Risk}(A) = c - P(A|D)$ and risk minimization yields the MAP estimate as the optimum Bayesian decision. Other well-known choices for the loss function include the quadratic error and the absolute error; i.e., assuming scalar S, A, we can select $L(S, A) = |S - A|^b$ with $b = 1, 2$. For $b = 2$ the risk is the Mean Square Error (MSE) and the optimum solution becomes the *mean* of the posterior distribution (i.e., the conditional mean given the data), whereas for $b = 1$ the risk is the Mean Absolute Error (MAE) and the optimum solution is the *median* of the distribution.

Returning to the view of perception as the process of inverting the world-to-signal mapping, this is generally an ill-posed problem. Thus, we need constraints to make it *well-posed*, i.e., to have a unique solution and the solution to depend continuously on the data. This approach is partially inspired by Tikhonov's *regularization theory* [113], which, to make inverse problems well-posed, proposes that we introduce some *constraints* by forcing the solution to lie in a subspace of the solution space where the problem is well-defined. For multimodal sensory perception, constraints can be of the following three types [17]: (i) Physical constraints, which stem from physical laws governing the multimodal world and are universally valid; (ii) Natural constraints that depend on the specific tasks (e.g., the smoothness constraints used in computer vision); and (iii) Artificial constraints that are imposed at some higher

cognitive level. Two important problems are to determine which constraints to use and how to embed them into the information processing algorithms.

An intuitive approach to incorporate constraints is the Bayesian formalism, where the plausibilities of different solutions are quantified by probabilities based on stochastic sensor models for the signal formation and prior expectations of the world state; the latter are influenced by previous measurements (as in active vision or ecological optics) and by the constraints we impose on the system. Then, as true solution we choose the one with the highest probability. As described in [57], in psychophysics, *ideal observers* are considered the theoretical observers who use Bayesian inference to make optimal interpretations. Usually 'optimality' is the MAP criterion since this allows an ideal observer to take into consideration both prior knowledge about the world's audiovisual structure as well as knowledge about the audiovisual signal formation by the sensors. In psychophysical tests, the ideal observer's optimum performance is a useful reference that is compared with the performance of a human observer.

Another convenient way to embed constraints for making the inversion of the world-signal mapping well-posed is via the *energy minimization* approach, which has become quite popular in computer vision and is closely related to regularization [47, 86, 37, 67, 76, 123]. Here the optimum audiovisual scene state $\hat{S}$ is found as the minimizer of the energy functional

$$E(S;D) = E_{data}(S;D) + E_{smooth}(S) \tag{1.6}$$

where the energy term E_{data} expresses a norm of the deviation of the scene S from the data D, whereas the term E_{smooth} measures the non-smoothness of S and hence imposes regularization constraints on the solution. The minimization of E is equivalent to maximizing the following Gibbs probability distribution for the posterior

$$P(S|D) = \frac{\exp[-E(S;D)]}{Z} \tag{1.7}$$

where Z is a normalization factor (the partition function). In this case, solutions that are consistent with the constraints correspond to lower energy states, which are chosen by the minimization process. We can see several correspondences between the energy and the Bayesian approach if we take logarithms on both sides of the Bayes formula (1.1): the data-fitting error energy E_{data} corresponds to the log likelihood $-P(D|S)$ and the regularization energy E_{smooth} is the log prior $-P(S)$. Obviously, the Bayesian formulation subsumes the energy minimization approach and offers a richer interpretation using statistical tools. For example, using the popular quadratic energy functions corresponds to assuming Gaussian distributions. Further, regularization problems that use multiple energy constraint terms correspond to a Bayesian prior that is the product of the individual priors assuming independent sources.

1.2.3 Examples of Multicue or Multimodal Perception Research

In this section we outline the main findings from a few selected works on multimodal perception. The particular papers were selected either because they have become classic in the field, like [71] that presents an archetypal example of (i) the brain combining sound and vision, or because they represent different viewpoints of research in multimodal perception that are directly related to this chapter's scope, like (ii) promoting individual visual cue features in weak fusion to facilitate their integration [62]; (iii) exploring the difference between intramodal versus intermodal fusion [46]; (iv) integrating audio and visual modalities to improve spatial localization [6, 28, 117]; (v) investigating the temporal segmentation of multimodal time-evolving scenes into perceptual events [125]; and (vi) using gestalt principles to group audio and visual modalities [72].

McGurk effect: Hearing Lips and Seeing Voices

McGurk and MacDonald's 1976 paper [71] is a classic on human sensory integration. The McGurk effect is elicited when a listener's perceptual report of a heard syllable is influenced by the sight of the speaker mouthing a different syllable, inducing the report of another syllable. This effect can be explained by assuming that the finally perceived syllable is the one mostly compatible with both conflicting stimuli. Specifically, by synchronously combining the original vocalizations and lip movements, dubbed videos of the type [ba-audio/ga-visual] and [ga-audio/ba-visual] were shown to subjects under audiovisual and audio-only conditions. The audiovisual presentations of speech caused two distinct types of responses: **'Fusion'** where the information from the two modalities is transformed into something new with an element not presented in either modality, and **'Combination'** where a composite is formed comprising relatively unmodified elements from each modality. To [ba-audio/ga-visual] presentations, almost all adults gave fused responses [da]. To its complement, [ga-audio/ba-visual], more than half gave combination responses like [gabga]. The effect is generalizable to other stop consonants.

To explain the [ba-audio/ga-visual] case, first note that /ba/ sounds somewhat similar to /da/. Further, there is some visual similarity between (the articulation of) the back consonant in /ga/ and the middle consonant in /da/, whereas there is no such similarity between /ga/ and the front consonant in /ba/. If we assume that, when presented with the two modalities, perceivers attempt to interpret an event by searching for something that has the most common features or best matches with both modalities, then the unifying percept is /da/. However, in a [ga-audio/ba-visual] presentation, the modalities share no common features and hence are in conflict. The listener cannot decide between the two modalities and oscillates between them, hearing various combinations [bagba, gabga, baga, gaba].

The main conclusions from [71] include: (1) Speech perception seems to also take into consideration the visual information. Audio-only theories of speech are inadequate to explain the above phenomena. (2) Audiovisual presentations of speech create fusion or combination of modalities. (3) One possible explanation of the two response types is that a human attempts to find common information in both modalities and achieve a unifying percept.

The above paper has inspired much work in exploring and reaffirming the bimodality of speech perception. An interesting issue is that of *complementarity*, stated in [69] as: "Not only audible and visible speech provide two independent sources of information, but each also provides strong information where the other is weak." For example, /bi/ and /pi/ are visually indistinguishable but can be distinguished acoustically based on features such as voice onset time. In contrast, /mi/ and /ni/ sound very similar but differ visually in the place of articulation. In both cases, audiovisual speech can aid detecting the differences.

Modeling Depth Cue Combination using Modified Weak Fusion

Landy et al. [62], taking the application of scene depth reconstruction from various visual cues as a showcase, examined in detail how the different cues can be combined to yield a fused final result. For scene depth reconstruction, the different cues examined are motion parallax (with known camera egomotion), stereo, kinetic depth effect, texture, and shading. These alternative cues are quite different in nature: first of all, motion parallax can provide absolute depth estimates, whereas the other cues provide stereo measurements up to some unknown parameters, for example up to the unknown viewing distance parameter. Inter-cue interaction can be employed then to resolve these parameters and make the measurements from different cues commensurate, in a process the authors call *cue promotion*. After cue promotion, all measurements are on the same scale and in common units. Then, promoted cues can be directly fused in a modified weak fusion scheme. (The simple weak fusion scheme of [17] does independent processing of each cue followed by a weighted averaging; see also Section 1.3.)

Beyond cue promotion, the authors introduce in their modified weak fusion scheme two further important enhancements relative to [17]: First, they underline the importance of *dynamic cue weighting*, in response to the spatial (presense or absense of certain cues in the scene) and temporal relative reliability of each cue. Second, they highlight the issue of *robustness* in combining the different cues, proposing that an explicit mechanism should be present for outlier detection and down-weighting. The three constituents, namely cue promotion, dynamic weighting, and robustness constitute the main aspects of what they term the **modified weak fusion** scheme. This scheme generalizes the weak fusion scheme of [17] in the sense that it allows limited interactions between the different cues (most notably for cue promotion), while at the same

time being modular and clearly more easy to verify than arbitrary strong fusion schemes. The authors give a methodology to assess the validity of the proposed fusion mechanism, as well as sufficient physiological experimental results in defense of their scheme.

Intramodal versus Intermodal Fusion of Sensory Information

Hillis et al. [46] explored human perception's capabilities for multimodal fusion to improve estimation of object properties (such as shape surface perception) both in an *intramodal* (within-senses) scenario of integrating the two visual cues of stereopsis-disparity and texture as well as in an *intermodal* (between-senses) scenario of integrating the two senses of vision and haptics. As optimal cue integration, they used a simple weak fusion [17, 124] where (under the Gaussian noise assumption) the Maximum Likelihood Estimate (MLE) becomes a linear weighted averaging with cue weights being inversely proportional to the variance of each cue noise. By performing psychophysical experiments and comparing the three cases of having (i) only single-cue estimators, (ii) only fused estimators (MLE), and (iii) both single-cue and fused estimators, they concluded to the following: Fusing cues (and losing information about the individual estimates) is more likely in the intramodal (disparity-texture) case than in the visual-haptic case. In the intermodal case, there may be natural circumstances where it is not beneficial to combine the two modalities (e.g., when one touches one object while looking at another).

Integration of Visual and Auditory Information for Spatial Localization

There is ample evidence that the human brain integrates multiple sensory modalities to accomplish various inference tasks such as spatial localization. In general, this integration improves performance. However, it may also lead to illusionary perception phenomena such as the "ventriloquist effect", where the movement of a dummy's mouth alters the perceived location of the ventriloquist's voice and hence creates a *localization bias*. Such phenomena are caused when there exist appropriate spatial and temporal disparities between the visual and auditory modalities. Experimental evidence [11, 117] has shown that the cross-modal localization bias decreases with increasing spatial and/or temporal disparity in the two stimuli.

Driver [28] explored variations of the ventriloquist illusion in the presence of a single visual and two auditory stimuli (target and distractor messages). He found that, under certain spatial combinations of the stimuli, the ventriloquist effect can actually help to enhance selective listening. Specifically, the simultaneous presence of a human face visually mouthing the target message and a mislocated sound source generating target and distractor messages can create an apparent visual source of the target sounds. Thus, before attentional

selection is completed, ventriloquism causes a cross-modal matching that spatially shifts the target sounds to a virtual instead of the actual location. This enhances the selective listening of the target message by focusing attention on the virtual source.

In several controlled experiments on integration of auditory and visual stimuli with spatio-temporal disparities, Wallace et al. [117] explored the relationship between two important aspects of multisensory integration, the perceptual unification of the two stimuli and the dependence of localization bias on their spatio-temporal disparities. They found that: (i) "regardless of their disparity, whenever the auditory and visual stimuli were perceived as unified, they were localized at or very near the light. In contrast, when the stimuli were perceived as not unified, auditory localization was often biased away from the visual stimulus"; (ii) "localization bias was a reliable, significant predictor of whether perceptual unity would be reported".

Battaglia et al. [6] compared two theories of how human observers fuse the visual and auditory modalities for spatial localization. One theory predicts a nonlinear integration where the modality whose signal is more reliable dominates over the other in a winner-take-all scheme. This model is known as *visual capture*, because human perception is usually dominated by vision over hearing. A typical example is watching a film in a movie theater where the visual information comes from the screen whereas the auditory information (loudspeakers' sound) originates from the sides, but human observers usually perceive the sound origin as coincident with the location of the visual stimulus. The other theory advocates for a linear integration of the two modalities through a weighted *visual-auditory average*, which corresponds to a weak fusion scheme [17, 124]. The authors conducted experiments where human subjects heard broadband noise from several locations and viewed noisy versions of a random-dot stereogram of a Gaussian bump. In the multimodal phase of the experiments, a difference in location was introduced between the visual and auditory stimuli. The results indicate that, in low-noise conditions, the observers' judgement was usually dominated by vision only. But at large noise levels, the observers' judgement shifted to an averaging of the two modalities. The authors also investigated a hybrid approach and proposed a Bayesian model that combines both the linear weighted averaging and a prior expressing an overall bias to vision.

Temporal Segmentation of Videos into Perceptual Events by Human Brain

Given the major role that the temporal structure has in human perception, Zacks et al. [125] addressed the two fundamental questions of whether and how the human perceptual system performs temporal segmentation into perceptual events. Their experimental method involved participants who watched short videos of daily activities while brain images were acquired with fMRI scanning. All participants watched each video in three corresponding modes:

naive passive viewing, intentional viewing seeking active segmentation into coarse time units and active segmentation into fine time units. The hierarchy between segmentation into coarse events and fine segmentation into subevents is conceptually similar to the spatial vision task of segmentation into objects and subparts. The main authors' conclusions are that there is significant and detectable neural activity in the human brain during both intentional and passive viewing, and this activity occurs around the perceptual event boundaries. Further, there is a hierarchical structure between the coarse and fine levels of segmentation, which are aligned. Finally, the segmented events are well correlated with environmentally meaningful parts of the video activity. Regarding this chapter's scope, we emphasize that, one open research direction in the above area is to investigate the separate roles of the individual audio and visual modalities as well as their integrated multimodality in the above temporal percept segmentation.

Audiovisual Gestalts

Nowadays, gestalt psychology principles have become an inspiration for several approaches in computer vision. In a relatively new direction, Desolneux, Moisan and Morel [25] detect visual gestalts based on a perceptual principle due to Helmholtz by finding statistically meaningful parts to be grouped through searching for geometrical structures that largely deviate from randomness. This work was extended by Monaci and Vandergheynst [72] to detecting audiovisual events. The authors' work in [72] is motivated by strong evidences from previous computational (e.g., in [45, 22, 105]) and psychophysical experiments (e.g., in [28, 117]) that the integration of audiovisual information by humans is strongly assisted by the temporal synchrony of events in the two modalities. It uses the Gestalt psychology principle of time proximity to relate audiovisual fusion with gestalt detection where the audiovisual gestalts are co-occurrences of auditory and visual events. To develop a computational algorithm the authors used the Helmholtz principle, introduced in image analysis by [25]. By combining sequences of energy features for the audio and displacement features for the visual stream, they derive synchronization indicator sequences from which they detect statistically meaningful audiovisual events. Their results re-confirm the significance of the temporal proximity between audio and visual events for integrating the two modalities. Computational studies with a similar goal (i.e., the importance of audiovisual synchrony) have also been done in [45, 22, 105, 5, 52, 101].

1.3 Bayesian Formulation of Fusion

As discussed in Section 1.2.2, inference about the world state is the process through which the world-to-signal mapping is inverted. Since this inverse problem is generally ill-posed, we need constraints to make it well-posed.

Sensor fusion is needed to: (a) Reduce the dependence of a sensor on possibly invalid a priori (natural or artificial) constraints. (b) Reduce the uncertainty in parameter estimation due to errors in the sensor modeling of the world-to-signal mapping. (c) Reduce uncertainty due to measurement noise contaminating the noise free data. The book [17] dealt mainly with (a). An approach to incorporate uncertainty estimation into the fusion problem is proposed in Chapter **??** of this book.

Let S be the world state to be estimated, e.g., a vector of numeric or symbolic features-parameters representing properties of an audiovisual scene of the external world. Let D be the multimodal data representing signals collected by auditory sensors, visual sensors and other information sources (e.g., text), or D may represent features extracted from the raw signals. We write $D = (D_1, D_2, D_3, ...)$ to separate the modalities. For simplicity, in this section we assume only two modalities, aural and visual, producing data sets D_A and D_V, respectively; hence, $D = (D_A, D_V)$.

Clark and Yuille [17] have proposed a classification of fusion cases in terms of weak and strong coupling, which we shall call simply 'weak fusion' and 'strong fusion'. Next we summarize the main ideas for both cases in the Bayesian framework.

Weak Fusion

A clear case of *weak fusion* [17] occurs if the aural and visual information processing modules are independent and have their own likelihoods $P_A(D_A|S)$, $P_V(D_V|S)$ and priors $P_A(S)$ and $P_V(S)$ and produce two separate posterior distributions

$$\text{audio}: \quad P_A(S|D_A) = \frac{P_A(D_A|S)P_A(S)}{P_A(D_A)} \tag{1.8}$$

$$\text{vision}: \quad P_V(S|D_V) = \frac{P_V(D_V|S)P_V(S)}{P_V(D_V)} \tag{1.9}$$

See Fig. 1.1, where for simplicity we denote the audio data D_A by A and the video data D_V by V. Each monomodal posterior could give its own MAP estimate of the world scene:

$$\hat{S}_i = \underset{S}{\text{argmax}}\, P_i(D_i|S)P_i(S), \quad i \in \{A, V\}, \tag{1.10}$$

Afterwards, for fusion, the two separate estimates can be combined somehow to give a combined audiovisual estimate:

$$\hat{S}_{AV} = \text{fusion}(\hat{S}_A, \hat{S}_V) \tag{1.11}$$

where the fusion function can be either linear (e.g., a weighted average) or nonlinear (e.g., a max or min combination).

Consider the above case of weak fusion and suppose we wish to find the joint maximum a posteriori (MAP) estimate from the combined prior

$$P_{AV}(S|D_A, D_V) = P_A(S|D_A)P_V(S|D_V) = \frac{P_A(D_A|S)P_V(D_V|S)P_A(S)P_V(S)}{P_A(D_A)P_V(D_V)} \tag{1.12}$$

If the two monomodal MAP estimates $\hat{S}_A$ and $\hat{S}_V$ are close, then Yuille and Bülthoff [124] have shown that the joint MAP estimate is a *weighted average* of the two single monomodal MAP estimates $\hat{S}_A$ and $\hat{S}_V$. Specifically, assuming that the two single MAP estimates are close, expanding in Taylor series the log posterior around the point $\hat{S}_A \approx \hat{S}_V$ and keeping up to second order terms yields

$$\begin{aligned} \log P_{AV}(S|D_A, D_V) \approx \log P_A(\hat{S}_A|D_A) + \log P_V(\hat{S}_V|D_V) \\ -[w_{\mathrm{a}}(S-\hat{S}_A)^2 + w_{\mathrm{v}}(S-\hat{S}_V)^2]/2 \end{aligned} \tag{1.13}$$

where $w_i = -(d^2 \log P_i(S|D_i)/dS^2)(\hat{S}_i)$, $i \in \{\mathrm{a}, \mathrm{v}\}$. Maximization of (1.13) yields the following MAP estimate for the audiovisual problem:

$$\hat{S}_{AV} = \frac{w_{\mathrm{a}}\hat{S}_A + w_{\mathrm{v}}\hat{S}_V}{w_{\mathrm{a}} + w_{\mathrm{v}}} \tag{1.14}$$

Since $w_{\mathrm{a}}, w_{\mathrm{v}} > 0$, the combined MAP estimate (after weak fusion) is approximately a linear convex combination of the monomodal estimates.

For Gaussian distributions, the second-order expression in (1.13) becomes exact and the assumption about $\hat{S}_A \approx \hat{S}_V$ is not needed. In this case the weights are inversely proportional to each modality's variance σ_i^2, $i \in \{\mathrm{a}, \mathrm{v}\}$. Since $1/\sigma_i^2$ measures the reliability of each modality, the weights in the above scheme are proportional to each modality's reliability.

A similar situation, i.e., the combined optimum estimate to be the weighted average of monomodal estimates, would occur again in a weak fusion scheme where we wish to obtain maximum likelihood (ML) estimates. In this case too, the combined likelihood factors into two terms

$$P_{AV}(D_A, D_V|S) = P_A(D_A|S)P_V(D_V|S) \tag{1.15}$$

Then, by working as in the MAP case, a second-order Taylor series expansion of the logarithm of (1.15) would yield as optimum multimodal estimate again a weighted average as in (1.14), but the $\hat{S}$ symbols would mean ML estimates and the weights w_i would result from the values of the second derivative of the monomodal likelihoods at their maxima (which should be close).

Strong Fusion

In the previous weak fusion scheme, the two modalities are processed independently, their monomodal optimal (with respect to the MAP or ML criterion)

estimates are found, and then fusion occurs by combining the two single estimates into a multimodal estimate with a linear or nonlinear function.

In contrast, we have *strong fusion* [17] if we have a non-separable joint likelihood and a single prior; this gives as posterior

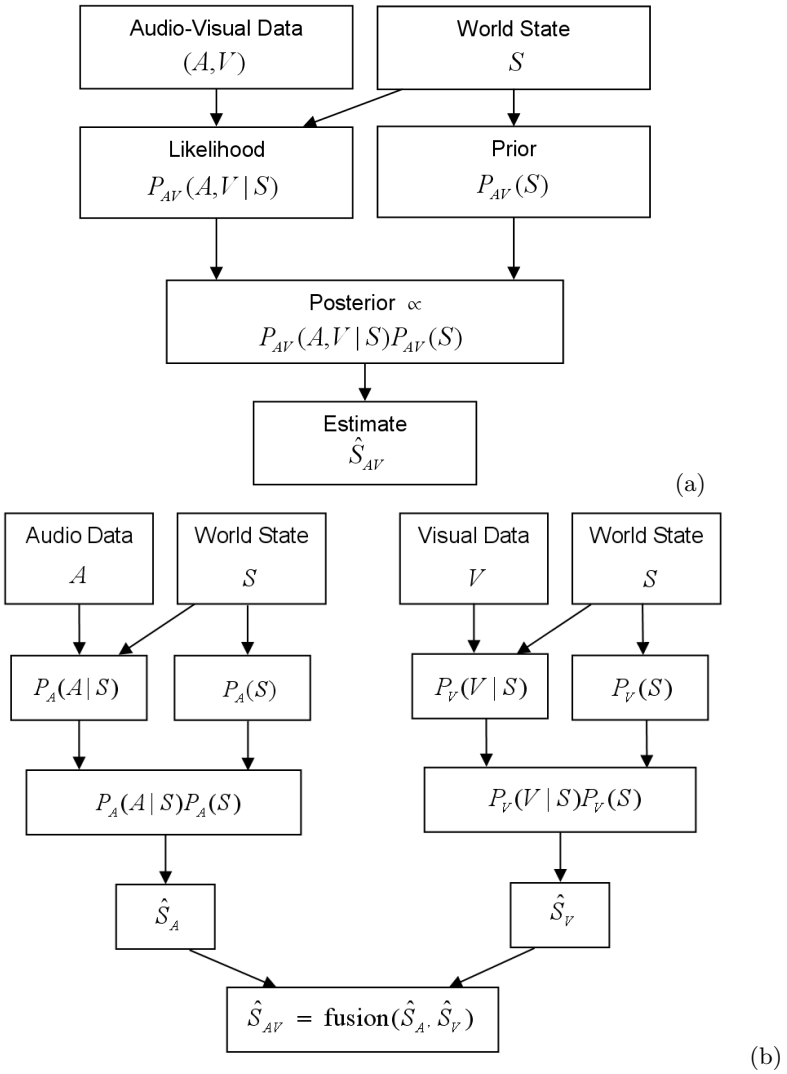

Fig. 1.1. Bayesian formulation of (a) Strong fusion and (b) Weak fusion schemes for two modalities, audio and vision.

$$P_{AV}(S|D_A, D_V) = \frac{P_{AV}(D_A, D_V|S)P_{AV}(S)}{P_{AV}(D_A, D_V)} \tag{1.16}$$

See Fig. 1.1 for diagrammatic illustration of strong fusion.

An *intermediate case* between weak and strong coupling is when the likelihood is separable and factors into two terms:

$$P_{AV}(S|D_A, D_V) = \frac{P_A(D_A|S)P_V(D_V|S)P_{AV}(S)}{P_{AV}(D_A, D_V)} \tag{1.17}$$

In this case, if the two modalities have the same prior, i.e., $P_{AV}(S) = P_A(S) = P_V(S)$, then we have a weak fusion scheme; otherwise we get a strong fusion.

Multi-stream Weights

In several cross-modal integration schemes used in multimedia applications, such as audiovisual speech recognition, the two modalities (or cues) are simply fused in the statistical models used for recognition by raising the respective monomodal likelihoods to various exponents, called *stream weights*. Without loss of generality, let us assume that we have two streams, say audio and video, with data or features D_A, D_V. The posterior probability of a property of an audiovisual scene S to be estimated given the multi-stream data $D = (D_A, D_V)$ is given by (1.16). If the two streams are statistically independent, the joint likelihood $P_{AV}(D|S)$ and marginal distributions $P_{AV}(D)$ become separable and we obtain (1.12). If needed, we can relax the independence assumption and assume only that the joint likelihood $P_{AV}(D|S)$ factors into the two corresponding monomodal likelihoods, in which case we obtain (1.17). The first case corresponds to simple weak fusion, whereas the second case is the aforementioned intermediate between weak and strong coupling. In both cases, raising each monomodal likelihood to a positive power, as usually done in multi-stream feature combination, creates a modified posterior-like function

$$B(S|D_A, D_V) = [P_A(D_A|S)]^{q_1}[P_V(D_V|S)]^{q_2}\frac{P(S)}{P(D)} \tag{1.18}$$

This may not even be a proper probability. Another artifact is the following: Since the rational numbers are dense in the set of reals, we can assume that the weights are rationals $q_i = n_i/n$, $i = 1, 2$, where n, n_1, n_2 are positive integers. If we ignore the common denominator n the integer stream weights correspond to replacing the product $P_A P_V$ of the marginal likelihoods with the power-weighted version $P_A^{n_1} P_V^{n_2}$. This corresponds to augmenting the multi-stream data (D_A, D_V) by replacing the ith stream with its n_i-fold repetition and treating the repetitions as independent. This repetition builds artificial correlations among subsets of the augmented data and may destroy any assumed independence of the separate streams. A better approach than power-raising the stream probabilities is proposed in this book's Chapter **??**.

1.4 Monomodal Features

For multimodal integration, a proper representation of each single modality is very crucial. Multimedia description standards such as the MPEG-7 particularly emphasize the role of monomodal descriptors [65]. Two main types of elements that MPEG-7 uses to describe and manage audio-visual content are descriptors and descriptor schemes. The *Descriptors* convey information about low-level audio features (e.g. temporal and spectral envelopes, pitch, energy, and descriptors about musical timbre, spoken content and melody contour) or visual features (e.g. color, edges, shape, texture, motion), as well as attributes of audio-visual content (e.g. location, time, quality). The *Description Schemes* convey information about high-level features such as audio segments and events or visual regions and objects; they consist of a set descriptors and their relationship for a particular task or application, arranged in a tree structure. The domain of descriptor schemes categorizes them into three types: audio, visual, and multimedia. The latter combine content descriptions of audio, visual, and possibly text data. Overviews of the MPEG-7 audio, visual, and multimodal descriptors and descriptor schemes can be found respectively in [89], [104], [97].

Next we discuss some of the most popular techniques to extract features from the audio, visual, and text data for multimodal applications.

1.4.1 Audio Features

Information carried by the audio modality is certainly polymorphous and multi-level. Thus, choosing a proper audio representation is not always straightforward and depends on the specific application. For multimodal integration, the need to also compactly capture properties that are complementary to the other modalities poses additional requirements. This has led to the utilization of simpler and more focused audio feature extraction schemes in many multimodal scenarios. Alternatively, the audio frontend is adopted 'as-is' from the corresponding audio-only application, e.g., speaker identification.

From a different viewpoint, audio descriptions in the multimodal context, inspired from single modality approaches or not, can be either *generic* or *specific* [68]. Specific refers to high-level representations of audio content, as for example obtained by applying speech recognition or audio segmentation and classification; knowledge of the audio class, e.g., music, noise, speech, or of the words that have been uttered in the case when the audio contains speech, can be used in a successive multimodal analysis and fusion scheme as sketched in Section 1.5.1. Such representations can be very useful in multimedia applications and are usually devised via the employment of advanced pattern recognition techniques. The focus of the current discussion is however mainly on the generic, low-level audio features as they are extracted by a signal-processing front-end. These may be further categorized into spectral or

temporal features or alternatively, according to [119], into short or long term audio representations.

Short-Term Features

Short-term features are normally extracted at a high rate and correspond to short audio segments, typically referred to as frames. They are advantageous in the sense that they allow the description of non-stationary audio signals whose properties vary quickly, e.g., speech. They cannot represent however long-term properties, as for example speech prosody. Probably the most widely used frame-level features are the short-term energy and variants as well as the zero-crossing rate [119, 122]. Their estimation is performed in the time-domain and when combined they can provide valuable information for speech/silence discrimination. Pitch, on the other hand, is extracted either by time or frequency analysis. It is the fundamental frequency of an audio waveform and it is very informative for speech and harmonic music. In speech, it mainly depends on the speaker's gender while in music it is influenced by the strongest note being played. In [70] the pitch is used, along with other features, for automatic meeting action analysis.

In applications such as audiovisual automatic speech recognition or audiovisual speaker identification, spectrum related representations are commonly preferred. This is justified by the relative success of the spectral feature sets in the corresponding audio-only based applications. Log-Mel filter-bank energies, Mel-frequency cepstral coefficients (MFCCs), linear prediction cepstral coefficients (LPCCs) [90], or perceptual linear prediction coefficients (PLP) are possible variants that have been successfully applied in various multimodal contexts [1, 30]. They practically provide a compact representation of smooth spectral information and their extraction is quite straightforward. For the most common ones, namely the MFCCs, the extraction process involves filtering the signal with a specially designed filterbank that comprises triangular or more elaborate filters properly localized in the frequency domain. The MFCCs are extracted as the first few Discrete Cosine Transform (DCT) coefficients of the log-energies of the signals at the filterbank output. Their efficacy is demonstrated in the context of audiovisual speech recognition in Chapter **??**. Usually, to capture speech dynamics, these features are also accompanied by their first and second derivatives. These derivatives are approximated using information from neighboring frames as well and so they would be more appropriately characterized as long-term features.

Long-Term Features

Long-term feature estimation is based on longer audio segments, usually comprising multiple frames. In a sense, long-term features capture variations of the short-term ones and may be more closely related to audio semantic content. Many such features were originally applied for audio analysis in single-stream

approaches and were or can be further customized for multimedia applications [126, 85]. Examples include various statistics of the short-time energy and the Zero Crossing Rate (ZCR), such as their average value or standard deviation. These indicate temporal changes of the corresponding quantities which in turn facilitate audio segmentation or classification in various classes, e.g., sports or news clips [119]; a sports clip would have a smoother ZCR contour than a news clip, since it is characterized by an almost constant noise background while clean speech during news exhibits a widely and quickly varying ZCR.

Long-term features based on *pitch statistics* can be equally useful. Only voiced speech and music have smooth pitch and thus pitch variations can help detecting voiced and music frames within an audio segment [119]. In a similar manner, spectral variations in time can help determining between speech and music; speech is expected to have much faster varying spectral characteristics. Indeed, temporal stability, i.e., a feature measuring these variations has been successfully applied in this direction at the first stage of a broadcast news multimedia indexing framework [80]. It is estimated as the variance of the generalized cross-correlation of the spectra of adjacent frames. At a different level, the *speaking rate*, i.e., how fast speech is uttered, can also be important; it may change a lot depending on speech pragmatics, namely the goal the specific speech utterance serves in communication. Being much different in a monologue or a presentation than during a conversation, the speaking rate has been exploited in [70] for multimodal meeting analysis.

1.4.2 Visual Features

The visual modality is an extremely rich source of information. Although high-level visual scene understanding of arbitrary scenes is beyond the reach of current technology, visual information processing plays a key role in various application areas, especially in domains where the image/video content is structured according to well-defined rules. In particular, visual information processing has proven beneficial in commercially-interesting domains involving sports video, broadcast news, and movies data, where it has been utilized in conjunction with audio and text for automatic content analysis, indexing, summarization, and re-purposing, among others.

A wide range of visual features has been proposed in the literature to address the requirements of different multimodal applications. We can categorize different visual information representations into two broad classes, low and mid-level generic visual features on the one hand, and high-level application-specific visual features on the other hand. We discuss next representative approaches from both categories.

Low and Mid-Level Visual Features

In the first category, low and mid-level visual features have been used to characterize basic image and video properties, such as color, texture, and

motion. This class of features are broadly used as generic image descriptors, most notably for applications such as content-based image/video retrieval [48, 94, 119], see also Section 1.7.6, and variants of them have been included as descriptors in the visual part of the MPEG-7 standard [65, 104].

Among the generic visual features, *color* is perhaps the most widely used. Color is typically represented in one of standard color-spaces, such as the RGB or the perceptually better motivated HSV and $L^*u^*v^*$. The color content of an image or video frame is typically summarized in a low-dimensional color histogram signature, and color-based similarity metrics are defined in terms of these histograms. Since color is a per-pixel attribute, color representations can be computed very efficiently and are invariant to image scaling or view-point changes. However, color histograms can be significantly affected by illumination changes and, most importantly, completely discard geometric image information since they do not represent the spatial configuration of pixels. Color features are typically most appropriate either for applications in which their efficiency is crucial, such as real-time (audiovisual) face and hand tracking, or for scenarios in which a single color, such as grass' green in field games, dominates the scene and thus its detection safely identifies the video shot.

Another universal image attribute is its *texture*, roughly corresponding to a description of its periodic patterns, directional content, and structural/stochastic complexity. A popular way to describe texture is by the image response of a multi-scale, oriented transform, such as the Gabor or wavelet filterbanks. The textural content can then be characterized by the most dominant filterbank responses at each point, or by filterbank channel response moments across the whole image. Alternative textural representations, such as Markov-Random-Fields or co-occurrence matrices can also serve as powerful texture descriptors. Another successful recent advance in image texture modeling encompasses the mid-level vision class of distinctive image features popularized by the Scale-Invariant Feature Transform (SIFT) representation [64]. In the SIFT representation, a sparse set of salient scale-space image points is first selected, and then the image textural content around each salient point is described in a compact representation. This class of features has built-in scale invariance properties and has proven particularly effective for reliable image matching and higher-level object recognition.

The last large class of low and mid-level visual features represents *motion* in video, typically computed with block-matching or other optical flow techniques [4]. On the one hand, global scene motion can be used to recover camera attributes such translation, rotation, panning, and zooming, as well as detect scene changes. On the other hand, local object motion is often related to object saliency; motion features have thus been used widely as event detectors in surveillance and sports analysis applications. The down-side of motion-based features is that optical flow computation is a computationally demanding task.

We should note here that each of the low and mid-level visual feature classes described above is typically not adequate for describing by itself the semantic content of image and video. Thus, most applications typically utilize more than one visual cues in tandem (intramodal fusion), apart from combining them with audio or text information (intermodal fusion), or even allow the user participate in the processing loop, as in the relevance feedback approach of [95].

High-Level Visual Features

In certain domains high-level image and video understanding is indispensable and this is usually beyond the reach of the low and mid-level visual descriptors just described. Typical example applications are audiovisual speech/emotion recognition or face recognition, which all require high-level models for object analysis and recognition. We describe next representative models of object shape and appearance which are carefully tailored for the needs of high-level object understanding.

An important high-level visual attribute is the object's *shape*. Examples of generic shape representations are the region-based 'shape context' [8], which yields a histogram shape descriptor, and the classical contour-based Fourier shape descriptor which approximates a closed contour using the coefficients of a truncated Fourier series; variants thereof are included in the MPEG-7 standard [104]. A more powerful class of object-centric shape features are the parametric representations of [83] and [20]. Both these techniques capture shape deformation in a compact parametric shape model which is specifically tailored for a single class of objects. This object-specific shape model is derived either by a physics-based Finite Element Modeling (FEM) analysis in [83] or by a training procedure using a hand-annotated set as in the Active Shape Model (ASM) [20]. Given such a model, a target shape can then be described in terms of its first few eigen-modes or eigen-shapes, yielding a highly compact and specific representation. Such models have been used extensively in the past for applications such as tracking and feature extraction from human faces.

In parallel to shape, an important class of computer vision models is concerned with object *appearance* description. Popularized by the successful "eigen-face" model of [114], this class of models strives for accurate and compact representation of image appearance content. Such representations are typically learned from representative training images by means of principal component analysis or other unsupervised/supervised dimensionality reduction techniques. A significant recent advance in appearance models is the Active Appearance Model (AAM) [21] which combines the compact shape representation of the ASM and the PCA-based appearance modeling of eigenfaces in a new powerful blend, while at same time being amenable to efficient calculations. An application of the AAM model in audiovisual speech recognition is illustrated in the book's Chapter **??**.

1.4.3 Text Features

The basics of text description

Text is a major medium carrying semantic information. In this section, we focus on the textual features that can be used to describe the content of a document for applications such as information retrieval for example. The basic paradigm consists in associating with each document a descriptor, called *index* in this domain, composed of a set of words called indexing terms. Such terms can be chosen from a predefined list (e.g., in a thesaurus) – in this case the correspondence between a document and this list is not trivial and is often done manually – or directly from the text itself. The latter is the usual way to proceed when large collections are to be processed automatically, i.e., in most search engines on the web.

To develop such a system implies first to choose the terms that should be extracted. The first stage in this process transforms the text in a sequence of words or tokens. If this is not too difficult in English, the absence of white spaces in Chinese for example can make this first step a rather difficult one. Next, the indexing terms are to be chosen among all the extracted tokens. They should be discriminant, and thus not appear in all documents, but they should not be too specific: they must appear in several texts and be informative [100]. The set of unordered words obtained by this process is called a bag-of-words.

Many systems associate a weight with each of the indexing terms, in order to designate which terms are more important or more meaningful. Three criteria are used: the importance of the term within the document, the importance of the term within the document collection, and the size of the document [98]. The first factor corresponds to a local weight and is usually based on the *term frequency* in the document. The second one is global and is often chosen as the *inverse document frequency* or one of its variants. The last factor tries to correct the effects of the size of the document.

Finally, a representation model defines the way the terms should be used or interpreted and how the query index should be compared to the collection indexes. Classical families of such models are the *set-theoretic* models like the Boolean model, where the documents are represented as sets of terms and where the comparisons are done using basic set-theoretic operations, the *algebraic* models like Salton's vector space model [99], where the documents are represented as vectors and the similarity is expressed as a scalar number, and *probabilistic* models [109] where the retrieval problem is seen as a probabilistic inference problem, making use of tools like Bayes' theorem.

Natural language processing for enhanced descriptions

The basic tools presented above fail to represent all the details and subtleties of natural languages, and natural language processing methods have been

proposed in order to acquire linguistic information and to improve the performance of the description. These tools can work at various levels: at the morphological level, at the syntactic level or at the semantic level.

Morphology is concerned by the structure of the words, and explains the links between words like *transform*, *transforms*, and *transformation*. A basic idea, called lemmatization, is thus to replace all these words in the document index by the simplest one or the most basic one of the series: *goes* can be replaced by *go*, *bikes* by *bike*. According to experiments, such a technique allows to improve the precision and recall of an information retrieval system up to 20 %. A second technique pushes the idea further and replaces every word in the index by its stem, but the results are much dependent on the quality of the stemming algorithm used and on the language [3]: Swedish or Slovenian provide more convincing results than English.

The structure of the sentences and of the syntagms are the subject of syntax. Its use mainly consists in using syntagms as complex indexing terms. Although they present even more variations than simple terms, their use has been proven successful when they come in addition to the simple terms, directly in the same index [100] or in a separate index [61]. The gain in performance can reach 5 to 30 %.

At the semantic level, information about the meaning of words and relations between words can be taken into account. Possible relations are synonymy, hypernymy, or more complex relations like the one that links 'professor' with 'to teach' [18]. Such information can be used to expand the queries. Automatically extracted co-occurring words added to queries have been proved to improve the results [36], while the use of WordNet [33] leads to more deceiving results [116]. Another alternative is to use the semantic information in the index itself, by employing the meaning of the words as indexing criterion instead of viewing each word as a sequence of letters. Disambiguation is also an option [54]. In this case also, the use of WordNet does not clearly improve the results.

As a conclusion, if the basic ways to integrate linguistic information presented here have largely proven their relevance, really taking into account word meaning is still a challenge for which no universal technique is yet available. Tools developed for restricted domains (e.g., more restricted collections or specific languages) appear however very promising.

1.5 Models for Multimodal Data Integration

What tools can be used to analyze jointly the several media present in a document? Many authors tried to avoid developing ad-hoc methods for each new combination they encountered and relied on the classical techniques that were available in fields like data analysis, machine learning or signal processing. As a matter of fact, multimedia is yet another application domain for pattern recognition techniques.

Let us first categorize the tasks to be solved in four elementary problems. *Segmentation* aims at delimiting events. These events can be shots, scenes or sequences in an audiovisual stream. *Event detection* consists in finding predefined events in a document, such as advertisement, dialogues, and goals in soccer matches. *Structuring* is close to a complete segmentation of a document. Its goal is to provide the complete structure of a document, structure that can include some hierarchy (e.g., shots are gathered in scenes), or some classification (e.g., the various segments may be labeled). Finally, *classification* aims at providing labels to document parts. Of course, one major application of classification consists in associating more semantic labels to documents, but this leads to a very wide variety of problems, e.g., determining the language of a document, what is its genre, and what sport is shown.

These four categories have close links and many algorithms both segment and structure, or detect events and classify at the same time. It should be noticed that the first three tasks, as far as multimedia documents are concerned, deal most of the time with temporal documents and have to take this temporal dimension into account. On the other hand, the classification often arrives after other description steps and can be stated as a static problem. As written above, many temporal or dynamic algorithms also achieve classification tasks.

Several categorizations can be made of the various techniques. Section 1.5.1 introduces the distinction between early, intermediate, and late fusion. In [108], the authors present other typologies and separate statistical methods, ranging from rule-based techniques, or simultaneous methods where all the media are considered at the same time to methods where the media are processed one after the other. In Section 1.5.2 we describe appropriate representations, as well as classification tools for static modeling of multimodal data, while in Section 1.5.3 we describe tools suited for dynamic time-evolving modalities, including the Hidden Markov Model and its variants, as well as more general Dynamic Bayesian Networks.

1.5.1 Levels of Integration: Early, Intermediate, and Late Fusion Approaches

Integration of features extracted from diverse sources is not a trivial task. The two main problems encountered in this process are the following:

- A *decision* problem: what should be the final decision when the various media or sources of information provide contradictory data? Although the decision problem is common to all systems based on information fusion, it gets more difficult in the case of multimodal data because the different modalities are affected dissimilarly by environmental noise, and thus their relative reliability is time-varying.
- A *synchronization* problem, which is specific to multimodal integration of time-evolving data. Synchronization issues arise for two reasons. First, the natural representation granularity for heterogeneous modalities is different. For example, the elementary unit of video signal is the image frame,

typically sampled at 20-30 Hz, while audio features for speech recognition are usually extracted at 100 Hz, and the elements of text (words) are generated at roughly a 1 Hz rate. Second, the boundaries induced by a certain semantic event to different modalities are only loosely aligned. For example, applause (acoustic evidence) and score label update (textual evidence) typically lag scoring in sports, while visual evidence is concurrent to it.

One can generally classify the various approaches to multimodal integration into three main categories [42], depending on the stage that the involved streams are fused, namely early, intermediate and late integration techniques. In the early integration paradigm, corresponding to the strong fusion model of Section 1.3, we first concatenate all modality descriptors into a single multimodal feature vector; afterwards, processing proceeds by using conventional monomodal techniques. Late integration techniques, following the weak fusion model of Section 1.3, largely handle each modality independently using separate models; the corresponding partial results are subsequently combined to yield the final decision. While both early and late integration approaches build on established monomodal modules and are thus easily applicable, they cannot fully account for the loose synchronization and the fine interaction between the different modalities. Intermediate integration methods try to address this shortcoming by employing novel techniques specifically devised to handle multimodal data and properly account for multimodal interaction.

Early Integration

For early integration, it suffices to concatenate all monomodal features into a single aggregate multimodal descriptor, possibly compacted by a dimensionality reduction process. Since early integration corresponds to the strong fusion model of Section 1.3, it is theoretically the most powerful scheme for multimodal fusion. In practice, however, early integration schemes can only be effective if all individual modalities are synchronized. Moreover, early integration lacks flexibility due to its non-modular nature, and the whole system needs to be re-built in case the conditions affecting even a single constituent modality change. For example, in the case of audiovisual speech recognition based on early integration models, it is necessary to retrain all models for each acoustic noise condition.

Late Integration

In this approach, each modality is classified independently. Integration is done at the decision level and is usually based on heuristic rules. For example, audio and video streams are segmented and classified by two separate Hidden Markov Models. Dialogues are identified as segments where audio signal is mainly speech while visual information is an alternation of two views. The detection of such particular scenes is done by fusion of the decisions.

A particular instance of late integration techniques is based on the successive analysis approach. The principle of this scheme, as illustrated in video analysis applications, is the following: The audio or textual signal is employed in a first stage to detect interesting segments. Image analysis (tracking, spatial segmentation, edge/line/face detection) is then used in the regions previously detected to identify a particular event, or more simply to identify the video segment boundaries. In this first case, audio, or text, are used to restrict the temporal window where video analysis will be used. An implicit assumption of such a method is that interesting segment detection is faster with these modalities (applauds in the sound track or keywords in the textual stream). This constitutes the first stage of a prediction verification method, whose second stage is a verification and localization step done on the audio or on the visual stream. The use order of the various media may be inverted: in a first stage, visual features are used to detect interesting events. In a second stage, the state of excitement of the speaker or the public is measured to filter the most interesting shots. This process is no more a prediction/verification process, but the audio signal is used to order the visual segments by level of importance.

Intermediate Integration

Intermediate integration techniques lie in-between early and late integration methods and are specifically geared towards modeling multimodal time-evolving data. They achieve a good compromise between modularity and close intermodal interaction. Specifically, they are modular enough in the sense that varying environmental conditions affecting individual streams can be handled by treating each stream separately. Moreover, they allow modeling the loose synchronization of heterogeneous streams while preserving their natural correlation over time. This class of techniques has proved its potential in various application areas, such as audiovisual speech recognition presented in Chapter **??**. Various intermediate integration architectures for handling time-evolving modalities are discussed in Section 1.5.3.

1.5.2 Static Models for Multimodal Data

We first consider static models for processing multimodal information. These are designed for data that are static themselves, but can also often handle satisfactorily dynamic data on a frame-by-frame basis.

Modeling Interrelated Multimodal Events

Multimodal data stemming from a common cause often have strong interdependencies. In the case of two sets of continuous vector variables, $\mathbf{x}$ and $\mathbf{y}$,

canonical correlation analysis (CCA) provides a natural representation for analyzing their co-variability [66]. Similarly to the better-known principal component analysis (PCA), CCA reduces the dimensionality of the datasets, and thus produces more compact and parsimonious representations of them. However, unlike PCA, it is specifically designed so that the preserved subspaces of $\mathbf{x}$ and $\mathbf{y}$ are maximally correlated. Therefore CCA is especially suited for studying the interrelations between $\mathbf{x}$ and $\mathbf{y}$. In the case that $\mathbf{x}$ and $\mathbf{y}$ are Gaussian, one can prove that the subspaces yielded by CCA are also optimal in the sense that they maximally retain the mutual information between $\mathbf{x}$ and $\mathbf{y}$ [102]. Canonical correlation analysis is also related to linear discriminant analysis (LDA): similarly to LDA, CCA performs dimensionality reduction to $\mathbf{x}$ discriminatively; however the target variable $\mathbf{y}$ in CCA is continuous, whereas in LDA is discrete.

More specifically, in CCA we seek directions, $\mathbf{a}$ (in the $\mathbf{x}$ space) and $\mathbf{b}$ (in the $\mathbf{y}$ space), so that the projections of the data on the corresponding directions are maximally correlated, i.e. one maximizes with respect to $\mathbf{a}$ and $\mathbf{b}$ the correlation coefficient between the projected data $\mathbf{a}^T\mathbf{x}$ and $\mathbf{b}^T\mathbf{y}$

$$\rho(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^T R_{xy} \mathbf{b}}{\sqrt{\mathbf{a}^T R_{xx} \mathbf{a}} \sqrt{\mathbf{b}^T R_{yy} \mathbf{b}}}. \tag{1.19}$$

Having found the first such pair of *canonical correlation directions* $(\mathbf{a}_1, \mathbf{b}_1)$, along with the corresponding *canonical correlation coefficient* ρ_1, one continues iteratively to find another pair $(\mathbf{a}_2, \mathbf{b}_2)$ of vectors to maximize $\rho(\mathbf{a}, \mathbf{b})$, subject to $\mathbf{a}_1^T R_{xx} \mathbf{a}_2 = 0$ and $\mathbf{b}_1^T R_{yy} \mathbf{b}_2 = 0$; the analysis continues iteratively and one obtains up to $k = \text{rank}(R_{xy})$ direction pairs $(\mathbf{a}_i, \mathbf{b}_i)$ and CCA coefficients ρ_i, with $1 \geq \rho_1 \geq \ldots \geq \rho_k \geq 0$, which, in decreasing importance, capture the directions of co-variability of $\mathbf{x}$ and $\mathbf{y}$. For further information on CCA and algorithms for performing it, one is directed to [66].

Canonical correlation analysis and related ideas have proven fruitful in several multimodal fusion tasks. By searching in videos for the image areas that are maximally correlated with the audio one can spot audiovisual salient events. Applications include speaker localization and tracking, as well as video-assisted audio source separation (cocktail party effect) [45, 22, 52]. By maximizing the canonical correlation over a small shift window, one can also time-align asynchronous data streams, as demonstrated in [101]. Moreover, CCA is closely related to the optimal Wiener filter for linear regression [102]; this connection has been employed by [51] in recovering speech articulation from audiovisual data.

Classification of Static Multimodal Data

In the matter of classification with multimedia data, many techniques coming from classic pattern recognition can be used. Before discussing these techniques, several specificities should be outlined.

The algorithms can be employed at several levels. Their input can be descriptors or the output of monomodal classifiers [56]. Of course, the way the various media are mixed is important. The descriptors can be simply concatenated. When the various descriptors are of the same nature, the resulting vector can be reduced through a PCA or discriminant linear analysis. Concatenating descriptors of different nature like words with numeric descriptors is problematic since they correspond to very different kinds of distributions and metrics.

Simple Bayesian classifiers are a first class of possible classifiers for multimedia data. Support vector machines (SVMs) [115] are heavily used for at least three reasons. They are quite efficient in dealing with high dimensional data, they can manage non-linear separation boundaries, and, last but not least, free implementations are available which are quite simple to use[5].

Neural networks of different kinds, like multilayer perceptrons, are also classical tools in the domain. Convolutional networks have been used for face detection and proven to be well-suited tools for dealing directly with the signal [35]. Even if their use is more efficient in some cases, they remain difficult to apply, because of the complexity of the algorithms that are associated with them for training and because no widely available implementation exists (for the convolutional networks in particular).

Finally, Bayesian networks appear to be very flexible tools. Such networks allow to model any graph of dependency between random variables. The variables are simply represented as nodes in a graph where edges represent some dependence between two variables. One of the major advantages of Bayesian networks comes from the possibility to learn the structure of the network directly from data, e.g., using the K2 algorithm [19]. Of course, if many variables are to be taken into account and no hint is given to the algorithm, this requires lots of training data and the complexity becomes very high. This is a major difference with Markov models where the structure has to be *a priori* defined.

1.5.3 Models for Dynamic Data: Integration of Asynchronous Time-Evolving Modalities

In the case of dynamic data, two additional difficulties appear: The various data streams can have different rates and can also lack precise synchronization. As an example, movies can have 24, 25 or 30 images per second when sound frames have a rate of 16 kHz or 48 kHz and speech corresponds to four syllables per second. It is also clear that TV and radio commentators usually describe events that have already passed, for example, in live sport programs. Even if the interval between the event and its comments is rather small for human perception, it will be translated in terms of dozens of image frames and hundreds of sound frames. Choosing what part of each stream should be

[5] A list is provided on the Wikipedia webpage on SVMs.

considered at a given instant is thus quite a complex problem. The two basic formalisms used in the domain are Markov models and Bayesian networks, the former being a particular case of the latter.

The Principle of Markov Models

Markov models are composed of a graph of states linked by oriented edges. Each edge represents a possible transition between two states or the possibility to stay at the same state for several periods of time. Time being assumed to be discrete, at each instant, the process makes a transition from its actual state to another one and emits an observation. Such a system is parameterized by several sets of probabilities. A first set provides the probability distribution of the initial state from where the process starts at $t = 0$. The second set provides the state transition probabilities. The last set provides the probability distributions of the observations emitted at each state.

Many variants of the basic model have been developed [88, 118]. The probabilities are usually constant over time, but one could use varying probabilities. The basic Markov hypothesis states that the observation emitted and the transition only depends on the current state: Past is reduced to the current situation. But here also, a variant is possible where the past could be reduced to the knowledge of a given number of past instants. That is for example what happens when using n-gram models.

A Markov model is said to be hidden when the sequence of states is unknown. This is the case for example when a sequence of observations is known, and the issue is then to determine the most probable sequence of states $s = (s_1, s_2, \ldots, s_T)$ which could emit this sequence of known observations $(o_1, o_2, \ldots, o_T)$. Mathematically, the problem is thus to find the optimal sequence of states s^* such that:

$$s^* = \arg\max_{s_{1:T}} (\log P(o_{1:T}|s_{1:T}) + \log P(s_{1:T}))$$

The Viterbi algorithm [34] is used to solve this problem and provides a global optimum.

It should be noticed that the structure of the hidden Markov model should be defined *a priori*. The parameters can either be predefined or estimated by the Baum-Welch algorithm from example data.

The Principles of Bayesian Networks

As mentioned earlier, the Bayesian networks allow a set of random variables and their dependencies to be represented by an oriented acyclic graph. Each node corresponds to an observed or hidden variable and the edges represent the dependencies. An edge between a node A and a node B indicates that the variable represented by B depends upon the variable represented by A.

Of course, the absence of any edge between two nodes means these nodes are independent, or conditionally independent if they share a common parent.

The parameters of such a network are the conditional distributions of each node. These distributions provide the probabilities of each variable conditionally on its parent variables. The global joint probability of all variables can be computed:

$$P_\theta(x) = \prod_i P_{\theta_i}(x_i|\mathcal{A}_i)$$

where $\mathcal{A}_i$ is the set of parents of node i in the graph. The parameters θ_i correspond to the parameters of the conditional distribution at node i. The factorized form of this joint probability is the starting point of the algorithms that allow to learn the structure [19] and the parameters of such networks [55, 50].

Such Bayesian networks are called dynamic Bayesian networks when they represent a random process. Such a denomination is quite improper in fact, but it is largely accepted by the community. Dynamic networks are in fact static, but they present a pattern repeated over time. On the other hand, the parameters are time independent. As a consequence, the same training algorithms can be used, but learning the structure of such a network becomes intractable in most situations.

Hidden Markov Models for Multimedia Processing

When several streams of observations are to be taken into account, Hidden Markov Models (HMMs) can be adapted. If the streams are synchronized and share the same rate, a first solution is to fuse the descriptors at each instant in order to create larger multimodal descriptors. Such a method is restricted to the fusion of descriptors of the same nature. Mixing words with numeric descriptors makes it difficult to define a metric between descriptors. Furthermore, the constraint on the rate often implies to align one of the streams on the other one (e.g., to reduce the audio information to one descriptor per visual shot) (see Fig. 1.2).

There are many Markov model variants for processing multimodal data; a unified presentation of the most popular architectures can be found in [79]. *Multistream* HMMs were introduced to process several streams, using one HMM per stream and by adding synchronization points [15, 44, 30] (see Fig 1.3). Between two such synchronizations, the two streams are assumed to be independent and are modeled by their own HMM. In this case the observations of the various streams are supposed to be independent (conditionally to the hidden process). At each synchronization point, the scores corresponding to each stream have to be combined.

Two extreme cases of multistream HMMs are the synchronous and the asynchronous models. In the former, the two Markov models have a shared state sequence, and can be considered as synchronized at every instant. In the

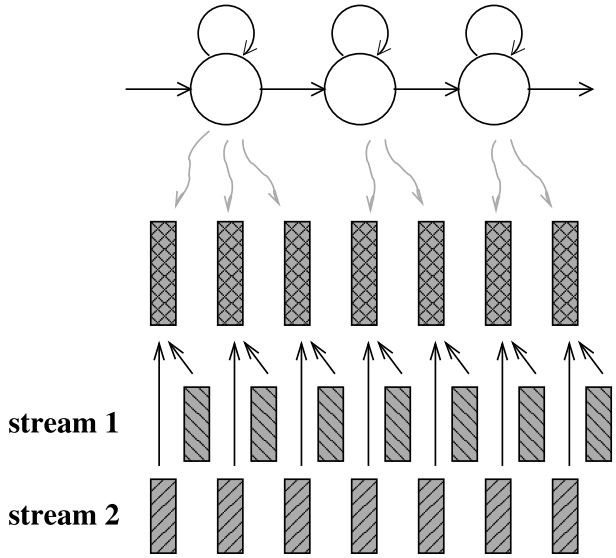

Fig. 1.2. Descriptor fusion with hidden Markov models. Grey arrows correspond to conditional probabilities and provide an exemple of alignment between states and observations. (Credits: G. Gravier)

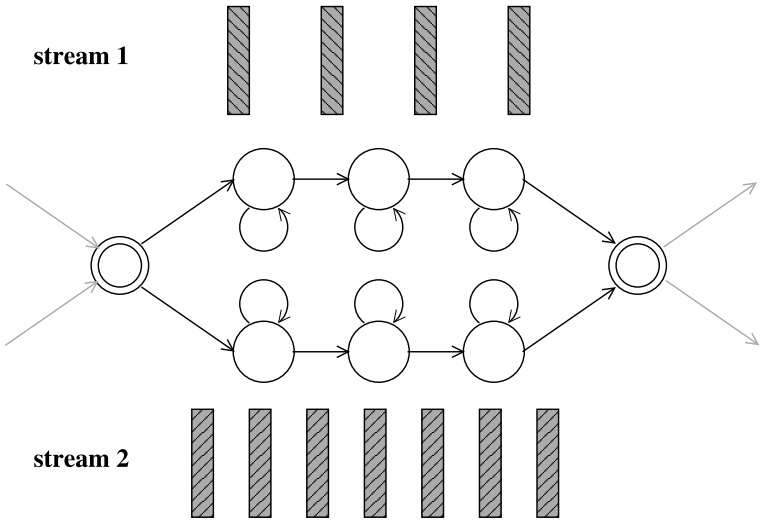

Fig. 1.3. Multistream hidden Markov model. The states represented by a double circle correspond to synchronization points. (Credits: G. Gravier)

latter, there is no synchronization (except at the beginning and at the end of the process) and the model is equivalent to a synchronous model in the product state space. As a consequence, such a model is often called a *product* model. More specifically, let us consider a pair of bimodal sequences $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$, each consisting of T (discrete or continuous) observation samples $\mathbf{y}^{(i)} = (y_1^{(i)}, y_2^{(i)}, \ldots, y_T^{(i)})$. Then, in the synchronous multistream HMM model the data are explained by a common hidden state sequence $\mathbf{x} = (x_1, x_2, \ldots, x_T)$, with x_t taking values in the single label set $\mathcal{L}$, yielding the overall probability

$$p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}|\mathbf{x}) = p(x_0) \prod_{t=1}^{T} p(x_t|x_{t-1}) p(y_t^{(1)}|x_t) p(y_t^{(2)}|x_t). \tag{1.20}$$

In the case of the asynchronous multistream HMM model, however, each modality has its own dedicated hidden state sequence $\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_T^{(i)})$, with $x_t^{(i)}$ taking values in the possibly separate label sets $\mathcal{L}^{(i)}$, yielding

$$p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = p(x_0^{(1)}, x_0^{(2)}) \cdot \prod_{t=1}^{T} p(x_t^{(1)}, x_t^{(2)}|x_{t-1}^{(1)}, x_{t-1}^{(2)}) p(y_t^{(1)}, y_t^{(2)}|x_t^{(1)}, x_t^{(2)}). \tag{1.21}$$

The resulting product HMM allows for state asynchrony, since at each time instance one can be at any combination of unimodal states.

The Bayesian network framework allows to represent easily other variants by introducing new possibilities of dependency between the states of the model. For example, Fig. 1.4 represents a *coupled* multistream model with a coupling between the chains associated to each stream. The associated observation sequence probability is

$$p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = p(x_0^{(1)}) p(x_0^{(2)}) \cdot \prod_{t=1}^{T} p(x_t^{(1)}|x_{t-1}^{(1)}, x_{t-1}^{(2)}) p(x_t^{(2)}|x_{t-1}^{(1)}, x_{t-1}^{(2)}) p(y_t^{(1)}|x_t^{(1)}) p(y_t^{(2)}|x_t^{(2)}). \tag{1.22}$$

Figure 1.5 represents yet another popular alternative, the *factorial* model. In this model, at a given instant, all hidden states depend upon all observations, yielding

$$p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = p(x_0^{(1)}) p(x_0^{(2)}) \cdot \prod_{t=1}^{T} p(x_t^{(1)}|x_{t-1}^{(1)}) p(x_t^{(2)}|x_{t-1}^{(2)}) p(y_t^{(1)}|x_t^{(1)}, x_t^{(2)}) p(y_t^{(2)}|x_t^{(1)}, x_t^{(2)}). \tag{1.23}$$

The relative merits of both multistream HMM variants, as well as the coupled and factorial HMM models, are examined by [79] in the context of audiovisual speech recognition.

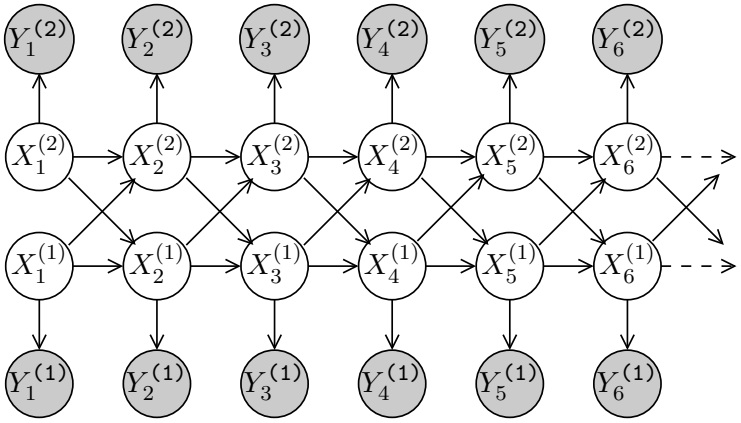

Fig. 1.4. Graphic representation of the coupled Markov model for two streams. States in grey correspond to observed states (the corresponding observations are not represented). (Credits: G. Gravier).

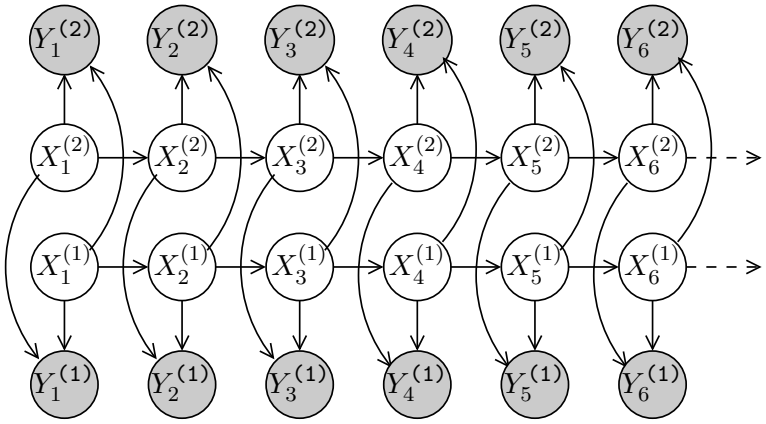

Fig. 1.5. Graphic representation of the factorial Markov model (Credits: G. Gravier).

Segment Models

In all the models presented so far, both Markov models and Bayesian networks associate a hidden variable to each observation. As a consequence, modeling a process that should stay for some time in a given state implies an exponential distribution for this duration. This is not always realistic.

Segment models [27, 81] are a variant of Markov models in which every state can be associated to several observations. Their number is modeled by an explicit duration model, which can be parametric or not. The use of such a model is for example, explained in the book's Chapter **??**.

1.6 Integrated Multimedia Content Analysis (Beyond Descriptors)

Although descriptor computation has attracted most of the attention of the video processing community, other aspects have also to be taken into account in order to derive complete systems. Several of these aspects are presented in this section, including metadata and the normalization problem, indexing techniques, and performance evaluation.

1.6.1 Metadata and Norms

The increasing number of digital photo and video collections raised the problem of describing them in a uniform way to allow an easier querying of these collections. One difficulty comes from the number of different communities that are concerned and have different habits and standards to describe their documents: documentalists used to manage libraries, the video community that developed the MPEG standards and wanted to enlarge their scope to include metadata through the MPEG-7 standard, the Web community which was confronted to the increasing number of images and videos and is working on the semantic web standards like RDF and OWL, some users like the American government defined their own system (the Dublin Core [23]). All these communities have their own standardization bodies like ITU for the telecommunication domain, ISO, IEC, the W3C, and this leads to a certain cacophony.

Another problem comes from the nature of the digital documents to be annotated. While books are material objects that do not change once published, even if a "book" may have several versions with differences between them, digital documents usually do not exist in a directly readable form but in compressed formats. They can be read through several software packages that provide different results which depend upon many factors like the screen used and the network bandwidth. Furthermore many versions of an original document can exist with various formats and resolutions. As a consequence, the concept of document has become quite fuzzy.

As far as digital videos are concerned, *metadata* can be separated into several categories. Some metadata describe the container of the document (e.g., name file, URL, compression format), some describe the physical aspects of the document when viewed (e.g., resolution, grey levels or color), others describe the content at various level: at low level with color histograms, at medium level with regions and spatial relations, at high level with faces / speakers and events. Another class of metadata is devoted to describing external elements: author, actors, how created, when broadcasted. Finally, some metadata describe the content from a human point of view, i.e., the story or the event. Such an annotation is not automatic because it requires some understanding of the document which is still impossible to achieve automatically with the current techniques.

The metadata normalization effort has led to different kinds of *norms*. First, some very general ones like the Dublin Core [23] which can be used for any digital document in fact. Although there are many variants, they all share a common basis and simplicity. Second, some attempts have been made to build a complete norm and led to MPEG-7 [65]. Such a norm suffers from several drawbacks: too general on the one hand, but not extensible on the other hand, not modular, based on a language that is not completely object-oriented and thus does not support inheritance although these properties were parts of the requirements [74]. MPEG-7 is a source of inspiration for many usages, but will probably not be used as such in practice.

Another norm very similar to MPEG-7 is nevertheless successful. TV-Anytime was developed at the same time but for a more focussed objective (the description of programs in TV streams) and, although based on the same concepts, was adopted and is used by many companies.

It is now accepted that conceiving a universal metadata norm is impossible and most standards plan to integrate and synchronize metadata with documents, but without specifying how these metadata should be written. This is, for example, the case of the MPEG-21 norm [73] or MXF [121].

1.6.2 Indexing Algorithms

Because in most libraries documents are described by words (authors, titles) and these words are sorted by alphabetic order, most persons fail to see the difference between a descriptor and an index which is a way to organize the metadata in order to retrieve easily the documents. As a matter of fact, the descriptors are used as indexes, and the alphabetic order is not seen as an external way to arrange them.

With images or sounds, the situation is different. Many descriptors are high dimensional numeric vectors subject to noise. Many of them cannot be compared directly, but through a distance or dissimilarity function and small differences may be considered as not significant. On the other hand, sorting such vectors in lexicographic order does not help to find the most similar vectors, whatever distance and search algorithm (ε-range or k-nearest neighbors searches) are used. Traditional database indexes of the B-tree family also fail to handle correctly such vectors for which all dimensions should be taken into account [2] at the same time.

A basic algorithm to solve this problem is the sequential and exhaustive search where all vectors are compared to the query vector. Such a simple algorithm has the major advantage to be absolutely linear in complexity. Many attempts have been made to improve this algorithm. A simple idea is to group the vectors in cells, and to be able to select the cells that should be read based on geometric properties.

Based on this idea, two main categories of algorithms have been proposed. Based on the seminal R-tree [40], some techniques called R+-tree [103], R*-tree [7], X-tree [10], SS-tree [120] and SR-tree [7] try to build the cells with

respect to the data distribution. All these cells are organized in a tree structure. On the other hand, other algorithms based on the KD-tree [9], like the K-D-B-Tree [91], the LSD-Tree or the LSD^h-Tree [43], build the cells by cutting the vector space in hyperrectangular regions. These regions are also organized in a tree. All these techniques appear to have a time complexity that grows exponentially with the size of the vector space, and none can be used in practice as soon as the space has more than 10 or 15 dimensions. This is one of the effects of the *dimensionality curse*.

New techniques are appearing that solve this problem. First, they implement an approximate search: Although finding close vectors to a query vector is not so long and difficult, proving that they are the nearest is time consuming. It was thus proposed to avoid this second stage. The tree structures have been abandoned in favor of linear structures (projection on random lines or on space filling curve, hash tables). Finally, distance computations can be replaced by rank aggregation techniques, following ideas developed in the OMEDRANK algorithm [32]. Grouping all these ideas into a single algorithm led to the PvS algorithm [63] which has a complexity close to constant in time and can be used in practice with billions of descriptors.

1.6.3 Performance Evaluation

Performance evaluation is a twofold concept. Firstly, it consists in assessing the quantitative and objective properties of a system, in a way that allows comparison with competing systems. Secondly, its goal is to verify to which extent a system fulfills the users' needs. These two kinds of evaluation give rise to very different techniques and methods.

Quantitative evaluation is now organized along a well established and recognized protocol through competitive campaigns. A set of experts firstly establish a test corpus and a learning corpus. For the first one, the experts establish a reference, i.e., they manually provide what is considered as a perfect result to which the systems will be compared. The second one is provided to the participants in order to develop, tune and test their systems. Secondly, a metric is chosen in order to compare the reference to the results that will be provided by each system. Finally, each participant runs their system on the test corpus and the results are compared to the reference using the chosen metrics. A workshop is often organized where all results are published and discussed. Many such campaigns are annual such that participants can improve their system and that new tasks can be addressed.

For example, the National Institute for Standards and Technology in the US organizes lots of such campaigns on various topics like information retrieval (TREC campaigns), machine translation, speech recognition, language recognition, speaker recognition, and video analysis.

This existing protocol nevertheless does not suppress the difficulties. A large community has to agree on one common task to be solved when lots of systems can solve slightly different problems. A metric should be agreed on.

Gathering the data and establishing the reference can be extremely difficult and expensive, because most multimedia data are copyrighted and because the manual annotation process for the reference is an extremely long and boring process (imagine you have to annotate every pixel of a long video!)

User evaluation is a completely different problem. Since it should involve users in real conditions, it can be achieved only on complete systems with interfaces and thus requires more development most of the time. Furthermore, the actions and reactions should be recorded and analyzed without perturbing its use of the system. Finally, questionnaires and spoken debriefings can complete the analysis.

Here also, the difficulties are numerous. Most computer scientists are not trained to manage such evaluations. Establishing any result often takes a lot of time and needs to involve many users to reduce any bias due the order of the data presented to each of the users and to the order of the tasks proposed to the user. Eventually, quantitative results can be obtained on only few simple questions, although the oral debriefing can bring more qualitative pieces of information.

1.7 Application Areas

Next we describe some indicative application areas in which multimodal integration techniques have proven particularly beneficial.

1.7.1 Audio-Visual Automatic Speech Recognition

Commercial *Automatic Speech Recognition* (ASR) systems are monomodal, i.e., only use features extracted from the audio signal to perform recognition. Although audio-only speech recognition is a mature technology [90], current monomodal ASR systems can work reliably only under rather constrained conditions, where restrictive assumptions regarding the amount of noise, the size of vocabulary, and the speaker's accents can be made. These shortcomings have seriously undermined the role of ASR as a pervasive *Human-Computer Interaction* (HCI) technology [82] and have delayed the adoption of speech recognition systems in new and demanding domains.

The important complementary role that visual information plays in human speech perception, as elucidated by the McGurk effect discussed in Section 1.2.3, has provided strong motivation for the speech recognition community to do research in exploiting visual information for speech recognition, thus enhancing ASR systems with speechreading capabilities [111, 88]. The key role of the visual modality is apparent in situations where the audio signal is either unavailable or severely degraded, as is the case of very noisy environments, where seeing the speaker's face is indispensable in recognizing what has been spoken. Research in this relatively new area has shown that multimodal ASR systems can perform better than their audio-only or visual-only

counterparts. The first such results where reported back in the early 80's by Petajan [84]. The potential of significant performance improvement of audio-visual ASR systems, combined with the fact that image capturing devices are getting cheaper, has increased the commercial interest in them.

The design of robust audiovisual ASR systems, which perform better than their audio-only analogues in all scenarios, poses new research challenges, most importantly:

- *Selection and robust extraction of visual speech features.* From the extremely high data rate of the raw video stream, one has to choose a small number of salient features which have good discriminatory power for speech recognition and can be extracted automatically, robustly and with low computational cost.
- *Optimal fusion of the audio and visual features.* Inference should be based on the heterogeneous pool of audio and visual features in a way that ensures that the combined audiovisual system outperforms its audio-only counterpart in practically all scenarios. This is definitely non-trivial, given that the audio and visual streams are only loosely synchronized, and the relative quality of audio and visual features can vary dramatically during a typical session.

These issues are discussed in detail in [88] and also in the book's Chapter **??**.

1.7.2 Sports Video Analysis, Indexing and Retrieval

Sport videos, as well as news reports, have motivated lots of research work due to their large number of viewers and possible applications [60]. The main challenge is to structure such videos in order to retrieve their structure or the main events they contain in order to navigate more easily, to index them or to derive new services from these videos.

Two categories of sport were especially studied: score oriented sports like tennis or volley-ball, which are organized depending on the score, and time oriented sports like soccer or rugby which are mainly organized in time periods with a variable number of events in each period. For the former case, the main goal is to recover the structure of the game and to evaluate the interest of each action [24]. For the latter case, the goal is to detect the interesting events [31].

Another usual problem is to separate the parts of video where the game is going on from all other instants like commercials, views on the public, and breaks. Sport video analysis thus combines some processing tools at various levels. Detecting the playing area is often the first step, but many other indices can be used according to the concerned sport: detecting players, detecting lines or areas on the playing area, detecting text, extracting the ball and tracking it. The sound track can be of great help, especially for event detection: applauds, pitch variation, keywords are usual cues. All these detectors are to be assembled in a global system. Stochastic models like HMMs or Bayesian

networks are a typical choice to fuse all the partial results in a global frame which allows recovering the structure of the video [53].

1.7.3 TV Structuring

TV structuring considers long and continuous TV streams of several days, weeks, or even months. In this case, the main goal is to compute an exact program guide, i.e., to segment the stream in smaller units and to characterize them by their start and end times and their title. These units are usually categorized into programs (e.g., weather forecast, news programs, movies) and non-programs (commercials, trailers, self-promotion of the TV channels, and sponsoring) TV structuring have mainly applications in the professional world for people working on TV archives, statistics or monitoring.

Two main methods have been proposed in the literature. A top-down approach [87] uses the regularity of program grids over years, and learns their structure from annotated data. The predicted grid is then compared to the stream to refine the detection of program separation. Such a method requires huge annotated data and was developed for TV archivers.

On the other hand, a bottom-up approach [78] tries to infer the stream structure directly from the stream itself. Most programs share no common information or structure that could be used to detect them. The segmentation thus starts by detecting the non-programs that have the common property to be heavily repeated in the stream. This can be achieved using a reference database or by directly comparing the stream with itself. Once the repetition are discovered and organized, the programs appear as the remaining segments. Their annotation can be done by comparing the stream with an Electronic Program Guide or the EIT tables associated with digital TV.

1.7.4 Multimedia Indexing of Broadcast News

Multimedia indexing of TV broadcast news programs is a very active application domain for the technologies of multimedia processing. There is significant interest in the potential of exploiting the vast amount of information carried over the TV networks on a daily basis. Exploitation in this context can be interpreted as the ability to efficiently organize, retrieve and reuse certain parts of the broadcasted information. This still poses various technological and scientific challenges and certainly effective multimodal integration of the involved audio, speech, text and video streams is one of the most important [16, 107, 80].

To classify broadcast news videos into various categories, in [16] they fuse low-level visual features such as color-histogram with audio class labels and high-level visual properties such as the number of faces appearing in the image. Classification is achieved via decision trees and the incorporation of multiple

modalities is shown to play a key role in the achieved performance improvement. A similar conclusion is drawn from the evaluation results for the multimedia indexing system presented in [80]. Initial story boundaries are localized using audio, visual and speech information and these are then fused in a weighted voting scheme to provide the final news story segmentation. A more elaborate fusion scheme, which unfolds at the semantic level, is proposed in [107]. Essentially, the concepts conveyed by the involved modalities, video and speech, are constrained to have certain relations between them. The 'Semantic Pathfinder', as the corresponding system is termed, exhibits quite promising properties in broadcast news indexing experiments.

1.7.5 Biometrics, Person Recognition

Automatic person recognition or identification processes have nowadays become indispensable in various transactions which involve human-machine interaction. Commonly, as for example at bank ATMs (Automated Teller Machines) or in transactions performed online, identification processes require issuing a certain token such as a card or just its number and then a password or a PIN (Personal Identification Number). To achieve increased security and naturalness, the utilization of physiological and behavioral characteristics such as the person's fingerprints, iris, voice, face or signature for identification, i.e., biometric recognition, is considered to be a much more promising alternative. However, fingerprint, iris and signature recognition, though quite reliable, involves high-cost sensors in many cases and is regarded as obtrusive. On the other hand, audio-only (voice) recognizers are cheap and quite user-friendly but vulnerable to microphone and acoustic environment changes. Similarly, visual-only (face) recognizers can be quite sensitive to lighting conditions and appearance changes. Integrated exploitation of two or more biometric modalities, appears to give the solution that satisfies requirements in each case [93, 1, 92, 106, 49, 77]. Audiovisual person recognizers for example significantly outperform the single-audio or visual recognizers in terms of reliability while at the same time feature low cost and non-obtrusiveness [1].

1.7.6 Image Retrieval and Photo-Libraries

The management and use of photo libraries has motivated a very large literature that is impossible to fully reference here. Several technologies have to be assembled in order to build a complete system. The choice of the components of the system depends upon the context of use of the system: Is the photo collection to be managed homogeneous or heterogeneous? Is the user a specialist or not? How the queries will be formulated?

Images cannot be compared or matched to the query directly in most applications. They have to be described or annotated first and comparison or matching will be performed on that description or annotation. This can be done using low level descriptors based on the image signal itself (color, texture

or shape descriptors) [14] or using only part of this signal (interest point, region descriptors) [64], using higher level processing tools (face detection, object recognition) [35], or keywords and text (coming from the image itself, associated to the image on a same web page or given by a human annotator). In association with these descriptors, a function must be defined in order to compare the image description with the query.

Another important question is the management of very large collections. When the descriptors are numeric, they are often represented as high dimensional vectors. Searching such vectors is a complex problem that is not solved by the use of database management systems [2].

Another key aspect is the user interface: this interface should allow the user to formulate its query and to see the results provided by the system. The user usually queries the system by presenting an image to the system (query by example) or by using words. Of course, the image descriptors used should be adapted to the queries; the matching between words and numeric descriptors remains a difficult challenge. Finally, the way the results are usually presented is a list of ordered images, although many works have also tried to develop other presentations.

1.7.7 Automated Meeting Analysis

Automated meeting analysis has lately come into the focus of interest in many diverse research fields, such as speech and speaker recognition, natural language processing and computer vision. The goal is to achieve systematic meeting indexing and structuring that would facilitate meeting information retrieval and browsing and would significantly favor remote meetings. In this direction, though speech is the predominant information carrying modality in this context, it has become clear that a meeting is essentially a sequence of multimodal human-human interaction processes and should be treated as such. Exploitation of video can help speaker and role identification in the meeting while the text of notes kept during the meeting may allow easier topic recognition and meeting segmentation. Proper consolidation of these modalities, i.e., video and text, in the analysis framework can lead to significant gains [70, 26, 13].

1.8 Conclusions and Future Directions

In this chapter we have surveyed some key ideas and results from research on cross-modal integration in multimedia analysis. We sampled problems from three major areas of research in multimedia: multimodal feature extraction, stochastic models for integrating dynamic multimodal data, and applications that benefit from cross-modal integration. In addition, we emphasized fusion of modalities or cues in various ways: explaining its weak- and strong-coupling

versions with a Bayesian formulation, classifying it at various levels of integration in conjunction with the stochastic classification models, and seeing it at work in various applications. As a useful supplement, we also reviewed a few ideas and results from human perception and its Bayesian formulation.

Some interesting future directions include the following:

Optimal Fusion: What is the best way to fuse multiple cues or modalities for various tasks and noise environments? Which should be the optimality criteria?

Fusing numeric and symbolic information: Multimodal approaches are now common for audio and video, for still images and text (or at least keywords). Mixing text or transcribed speech with video and audio is still a challenge, since it brings together numeric information coming from sound or images with symbolic information.

Investigate how the cross-modal integration algorithms *scale* and perform on large multimedia databases.

Cross-modal integration for performance improving in two *grand challenges:* (i) Natural access and high-level interaction with multimedia databases, and (ii) Detecting, recognizing and interpreting objects, events, and human behavior in multimedia videos by processing combined audio-video-text data.

Anthropocentric system: The interaction with the system and taking the human user into account are still open issues. And they are very important, since in most multimedia applications it is humans who will ultimately evaluate and use the system. Many aspects of human-computer interfaces are reviewed in this book's Chapter **??**.

Looking back at this chapter's journey, we attempted to take a few glimpses at a huge and fascinating field, that of multimedia understanding through cross-modal integration. We are still feeling that it is a very complex dynamic area. The understanding of each of the major sensor modalities, i.e., speech or vision, has not been "conquered" yet by science and technology, neither perceptually nor computationally. Imagine now their fusion! Nevertheless we must be brave and dare to keep researching this remarkable mapping from the combined audiovisual world to our multimodal percepts and inversely.

References

1. P. Aleksic and A. Katsaggelos, "Audio-visual biometrics," *Proc. IEEE*, vol. 11, pp. 2025–2044, 2006.
2. L. Amsaleg and P. Gros, "Content-based retrieval using local descriptors: Problems and issues from a database perspective," *Pattern Analysis & Applications*, vol. 2001, no. 4, pp. 108–124, 2001.
3. A. Arampatzis, T. Van Der Weide, C. Koster, and P. Van Bommel, *Linguistically Motivated Information Retrieval*. New York, New York, Etats-Unis: M. Dekker, 2000, vol. 69, pp. 201–222.
4. J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *Int'l J. of Comp. Vis.*, vol. 12, no. 1, pp. 43–77, 1994.
5. Z. Barzelay and Y. Y. Schechner, "Harmony in motion," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2007.
6. P. W. Battaglia, R. A. Jacobs, and R. N. Aslin, "Bayesian integration of visual and auditory signals for spatial localization," *J. of the Opt. Soc. Am. (A)*, vol. 20, no. 7, pp. 1391–1397, July 2003.
7. N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger, "The R*-tree: An efficient and robust access method for points and rectangles," in *ACM SIGMOD International Conference on Management of Data*, Atlantic City, New Jersey, Etats-Unis, 23-25 May 1990, pp. 322–331.
8. S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.
9. J. L. Bentley, "Multidimensional binary search trees in database applications," *IEEE Trans. Softw. Eng.*, vol. 5, no. 4, pp. 333–340, 1979.
10. S. Berchtold, D. A. Keim, and H.-P. Kriegel, "The X-tree : An index structure for high-dimensional data," in *22th International Conference on Very Large Data Bases*, Mumbai (Bombay), Inde, 3-6 Sep. 1996, pp. 28–39.
11. P. Bertelson and M. Radeau, "Cross-modal bias and perceptual fusion with auditory-visual spatial discordance," *Percept. Psychophysics*, vol. 29, pp. 578–584, 1981.
12. C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
13. M. M. Bouamrane and S. Luz, "Meeting browsing: State-of-the-art review," *Multimedia Systems*, vol. 12, pp. 439–457, 2007.

14. N. Boujemaa, S. Boughorbel, and C. Vertan, "Soft color signatures for image retrieval by content," in *Eusflat'2001*, vol. 2, 2001, pp. 394–401.
15. H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. Int'l Conf. on Spoken Language Processing*, 1996, pp. 426–429.
16. L. Chaisorn, T.-S. Chua, and C.-H. Lee, "A multi-modal approach to story segmentation for news video," *World Wide Web*, vol. 6, pp. 187–208, 2003.
17. J. J. Clark and A. L. Yuille, *Data Fusion for Sensory Information Processing*. Kluwer Academic Publ., 1990.
18. V. Claveau, P. Sbillot, C. Fabre, and P. Bouillon, "Learning semantic lexicons from a part-of-speech and semantically tagged corpus using inductive logic programming," *Journal of Machine Learning Research*, vol. 4, pp. 493–525, Aug. 2003.
19. G. F. Cooper and E. Herskovits, "A bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, pp. 309–347, 1992.
20. T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models – their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.
21. T. Cootes, G. Edwards, and T. C.J., "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, 2001.
22. T. Darrell, J. Fisher, P. Viola, and B. Freeman, "Audio-visual segmentation and the cocktail party effect," in *Proc. Int'l Conf. on Multimodal Interfaces*, 2000.
23. "Standard international iso/iec 15836,the dublin core metadata element set," Nov. 2003.
24. H. Denman, N. Rea, and A. Kokaram, "Content-based analysis for video from snooker broadcasts," *Computer Vision and Image Understanding*, vol. 92, pp. 141–306, November/December 2003.
25. A. Desolneux, L. Moisan, and J.-M. Morel, "Edge detection by Helmholtz principle," *J. Math. Imaging and Vision*, vol. 14, pp. 271–284, 2001.
26. A. Dielmann and S. Renals, "Automatic meeting segmentation using dynamic Bayesian networks," *IEEE Trans. Multimedia*, vol. 9, no. 1, pp. 25–36, January 2007.
27. V. V. Digalakis, "Segment-based stochastic models of spectral dynamics for continuous speech recognition," Ph.D. dissertation, Speech Processing and Interpretation Laboratory, Universit de Boston, 1992.
28. J. Driver, "Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading," *Nature*, vol. 381, pp. 66–68, May 1996.
29. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. J. Wiley & Sons, 2001.
30. S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.
31. A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. Image Process.*, vol. 12, no. 7, pp. 796–807, July 2003.
32. R. Fagin, R. Kumar, and D. Sivakumar, "Efficient similarity search and classification via rank aggregation," in *ACM SIGMOD International Conference on Management of Data*, San Diego, Californie, Etats-Unis, 9-12 Jun. 2003, pp. 301–312.

33. C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.
34. G. Forney, "The Viterbi algorithm," *Proc. IEEE*, vol. 61, no. 3, pp. 268–277, 1973.
35. C. Garcia and M. Delakis, "Convolutional face finder: A neural architecture for fast and robust face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1408–1423, 2004.
36. S. Gauch, J. Wang, and S. Rachakonda, "A Corpus Analysis Approach for Automatic Query Expansion and its Extension to Multiple Databases," *ACM Transactions on Information Systems*, vol. 17, no. 3, pp. 250–269, 1999.
37. S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, 1984.
38. E. B. Goldstein, *Sensation and Perception*. California: Wadsworth Publ. Co., 1984.
39. U. Grenander, *Elements of Pattern Theory*. The Jonhs Hopkins Univ. Press, 1996.
40. A. Guttman, "R-trees: A dynamic index structure for spatial searching," in *ACM SIGMOD International Conference on Management od Data*, Boston, Massachusetts, Etats-Unis, 18-21 Jun. 1984, pp. 47–57.
41. H. Helmholtz, *Physiological Optics, Vol.III: The Perceptions of Vision (J. P. Southall, Trans.)*. Rochester, NY: Optical Soc. Amer., 1910, 1925.
42. M. Hennecke, D. Stork, and K. Prasad, "Visionary speech: Looking ahead to practical speechreading systems," in *Speechreading by Humans and Machines*, D. Stork and M. Hennecke, Eds. Berlin, Germany: Springer, 1996, pp. 331–349.
43. A. Henrich, "The lsd^{h}-tree: An access structure for feature vectors," in *14th International Conference on Data Engineering*, Orlando, Floride, Etats-Unis, 23-27 Feb. 1998, pp. 362–369.
44. H. Hermansky, M. Pavel, and S. Tibrewala, "Towards ASR using partially corrupted speech," in *Proc. Int'l Conf. on Spoken Language Processing*, Oct. 1996, pp. 458–461.
45. J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," in *Proc. Advances in Neural Information Processing Systems*, 1999.
46. J. M. Hillis, M. O. Ernst, M. S. Banks, and M. S. Landy, "Combining sensory information: Mandatory fusion within, but not between, senses," *Science*, vol. 298, pp. 1627–1630, 2002.
47. B. K. Horn, *Robot Vision*. Cambridge, Massachusetts: MIT Press, 1986.
48. F. Idris and S. Panchanathan, "Review of image and video indexing techniques," *Journal of Visual Communication and Image Representation*, vol. 8, no. 2, pp. 146–166, 1997.
49. A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, no. 12, pp. 2270–2285, December 2005.
50. F. Jensen, S. Lauritzen, and K. Olsen, "Bayesian updating in recursive graphical models by local computations," *Computational Statistics Quarterly*, vol. 4, pp. 269–282, 1990.

51. A. Katsamanis, G. Papandreou, and P. Maragos, "Audiovisual-to-articulatory speech inversion using active appearance models for the face and hidden markov models for the dynamics," in *Proc. IEEE Int'l Conf. Acous., Speech, and Signal Processing*, 2008.
52. E. Kidron, Y. Y. Schechner, and M. Elad, "Cross-modal localization via sparsity," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1390–1404, Apr. 2007.
53. E. Kijak, G. Gravier, L. Oisel, and P. Gros, "Audiovisual integration for sport broadcast structuring," *Multimedia Tools and Applications*, vol. 30, pp. 289–312, 2006.
54. A. Kilgarriff and M. Palmer, "Special Issue on Senseval," *Computers and the Humanities*, vol. 34, no. 1/2, Apr. 2000.
55. J. Kim and J. Peal, "A computational model for causal and diagnostic reasoning in inference systems," in *Proc. Int'l Joint Conf. on Artificial Intel.*, 1983, pp. 190–193.
56. J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.
57. D. C. Knill, D. Kersten, and A. L. Yuille, *Perception as Bayesian Inference*. Cambridge Univ. Press, 1996, ch. Introduction: A Bayesian Formulation of Visual Perception, pp. 1–21.
58. K. Koffka, *Principles of Gestalt Psychology*. Routledge, 1935, 1999.
59. W. Köhler, *Gestalt Psychology*. New York: Liveright Publish. Corp., 1947, 1970.
60. A. Kokaram, N. Rea, R. Dahyot, A. M. Tekalp, P. Bouthemy, P. Gros, and I. Sezan, "Browsing sports video," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 47–58, March 2006.
61. W. Kraaij and R. Pohlmann, "Comparing the Effect of Syntactic vs. Statistical Phrase Indexing Strategies for Dutch," in *2nd European Conference on Research and Advanced Technology for Digital Libraries*, C. Nicolaou and C. Stephanides, Eds. Lecture Notes in Computer Science, Springer Verlag, 1998, vol. 1513, pp. 605–614.
62. M. S. Landy, L. T. Maloney, E. B. Johnston, and M. Young, "Measurement and modeling of depth cue combination: in defense of weak fusion," *Vision Research*, vol. 35, no. 3, pp. 389–412, 1995.
63. H. Lejsek, F. H. Asmundsson, B. Thor-Jonsson, and L. Amsaleg, "Scalability of local image descriptors: A comparative study," in *Proc. ACM Int'l Conference on Multimedia*, Oct. 2006.
64. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int'l J. of Comp. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
65. B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley, 2002.
66. K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*. Acad. Press, 1979.
67. J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solutions of ill-posed problems in computational vision," *J. of the Amer. Stat. Assoc.*, vol. 82, no. 37, pp. 76–89, March 1987.
68. J. Martinez, "Standards - mpeg-7 overview of mpeg-7 description tools, part 2," *IEEE Multimedia*, vol. 9, no. 3, pp. 83–93, Jul-Sep 2002.
69. D. Massaro and D. Stork, "Speech recognition and sensory integration," *American Scientist*, vol. 86, no. 3, pp. 236–244, 1998.

70. I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, pp. 305–317, 2005.
71. H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.
72. G. Monaci and P. Vandergheynst, "Audiovisual Gestalts," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition Workshop*. New York, NY: IEEE Computer Society, 2006, p. 200.
73. "Standard international iso/iec 21000 information technology – "multimedia framework"."
74. "MPEG-7 requirements document v.15, iso/iec jtc1/sc29/wg11, mpeg01/n4320," Jul. 2001.
75. D. Mumford, *Perception as Bayesian Inference*. Cambridge Univ. Press, 1996, ch. Pattern Theory: A unifying perspective, pp. 25–61.
76. D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems," *Commun. Pure & Appl. Math.*, vol. XLII, no. 4, 1989.
77. K. Nandakumar, Y. Chen, S. C. Dass, and A. K. Jain, "Likelihood ratio based biometric score fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, February 2008.
78. X. Naturel, G. Gravier, and P. Gros, "Fast structuring of large television streams using program guides," in *Proceedings of the 4th International Workshop on Adaptive Multimedia Retrieval (AMR), Geneva, Switzerland*, ser. Lecture Notes in Computer Science, vol. 4398, Jul. 2006, pp. 223–232.
79. A. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic bayesian networks for audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1–15, 2002.
80. K. Ohtsuki, K. Bessho, Y. Matsuo, S. Matsunaga, and Y. Hayashi, "Automatic multimedia indexing," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 69–78, 2006.
81. M. Ostendorf, "From HMMs to Segment Models," in *Automatic Speech and Speaker Recognition - Advanced Topics*. Kluwer Academic Publishers, 1996, ch. 8.
82. S. Oviatt, "Multimodal interfaces," in *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications*, J. Jacko and A. Sears, Eds. Mahwah, NJ, USA: Lawrence Erlbaum, 2003, pp. 286–304.
83. A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," *Int'l J. of Comp. Vis.*, vol. 18, no. 3, pp. 233–254, 1996.
84. E. Petajan, "Automatic lipreading to enhance speech recognition," Ph.D. dissertation, Univ. of Illinois, Urbana-Campaign, 1984.
85. A. Pikrakis, T. Giannakopoulos, and S. Theodoridis, "A dynamic programming approach to speech/music discrimination of radio recordings," in *Proc. European Signal Processing Conference*, 2007.
86. T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory," *Nature*, vol. 317, pp. 314–319, 1985.
87. J.-P. Poli and J. Carrive, "Modeling television schedules for television stream structuring, Singapour," in *Proceedings of ACM MultiMedia Modeling*, Jan. 2007, pp. 680–689.

88. G. Potamianos, C. Neti, G. Gravier, and A. Garg, "Automatic recognition of audio-visual speech: Recent progress and challenges," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.
89. S. Quackenbush and A. Lindsay, "Overview of MPEG-7 audio," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 725–729, 2001.
90. L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* NJ, USA: Prentice-Hall, 1993.
91. J. T. Robinson, "The K-D-B-tree: A search structure for large multidimensional dynamic indexes," in *ACM SIGMOD International Conference on Management of Data*, Ann Arbor, Michigan, Etats-Unis, 29 Apr. - 1 May 1981, pp. 10–18.
92. A. Ross and A. Jain, "Information fusion in biometrics," *Pattern Recognition Letters*, vol. 24, no. 13, pp. 2115–2125, September 2003.
93. A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of Multibiometrics.* Springer-Verlag, 2006.
94. Y. Rui, T. S. Huang, and S.-F. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *Journal of Visual Communication and Image Representation*, vol. 10, pp. 39–62, 1999.
95. Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool for interactive content-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 644–655, 1998.
96. B. Russell, *A History of Western Philosophy.* New York: Simon & Schuster, 1945.
97. P. Salembier and J. R. Smith, "MPEG-7 multimedia description schemes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 748–759, 2001.
98. G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," *Information Processing and Management*, vol. 24, no. 5, pp. 513–523, 1988.
99. G. Salton, *Automatic Text Processing.* Addison-Wesley, 1989.
100. G. Salton, C. Yang, and C. Yu, "A Theory of Term Importance in Automatic Text Analysis," *Journal of the American Society for Information Science*, vol. 26, no. 1, pp. 33–44, 1975.
101. M. Sargin, Y. Yemez, E. Erzin, and A. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1396–1403, Nov. 2007.
102. L. L. Scharf and J. K. Thomas, "Wiener filters in canonical coordinates for transform coding, filtering, and quantizing," vol. 46, no. 3, pp. 647–654, 1998.
103. T. K. Sellis, N. Roussopoulos, and C. Faloutsos, "The R+-tree: A dynamic index for multi-dimensional objects," in *Proc. Int'l Conf. on Very Large Data Bases*, Brighton, Royaume-Uni, 1-4 Sep. 1987, pp. 507–518.
104. T. Sikora, "The MPEG-7 visual standard for content description – an overview," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 696–702, 2001.
105. M. Slaney and M. Covell, "FaceSync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Proc. Advances in Neural Information Processing Systems*, 2001.
106. R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain, "Large scale evaluation of multimodal biometric authentication using state-of-the-art systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 450–455, March 2005.

107. C. G. M. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. W. M. Smeulders, "The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1678–1689, October 2006.
108. C. G. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, January 2005.
109. K. Spärck Jones, S. Walker, and S. E. Robertson, "A Probabilistic Model of Information Retrieval: Development and Comparative Experiments - Part 1 and 2," *Information Processing and Management*, vol. 36, no. 6, pp. 779–840, 2000.
110. R. J. Sternberg, *Cognitive Psychology*, 4th ed. Thomson Wadsworth, 2006.
111. D. Stork and M. Hennecke, Eds., *Speechreading by Humans and Machines*. Berlin, Germany: Springer, 1996.
112. S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 3rd ed. Acad. Press, 2006.
113. A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-posed Problems*. Washington DC: Winston & Sons, 1977.
114. M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 1991, pp. 586–591.
115. V. Vapnik, *Statistical Learning Theory*. New York: Wiley-Interscience, 1998.
116. E. Voorhees, "Using WORDNET for Text Retrieval," in WORDNET*: An Electronic Lexical Database*, C. Fellbaum, Ed. The MIT Press, 1998.
117. M. T. Wallace, G. E. Roberson, W. D. Hairston, B. E. Stein, J. W. Vaughan, and J. A. Schirillo, "Unifying multisensory signals across time and space," *Exp. Brain Research*, vol. 158, pp. 252–258, 2004.
118. F. Wang, Y.-F. Ma, H.-J. Zhang, and J.-T. Li, "A generic framework for semantic sports video analysis using dynamic Bayesian networks," in *International Multimedia Modelling Conference*, 2005.
119. Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis-using both audio and visual clues," *IEEE Signal Process. Mag.*, vol. 17, no. 6, pp. 12–36, Nov. 2000.
120. D. A. White and R. Jain, "Similarity indexing with the SS-tree," in *12th International Conference on Data Engineering*, 26 Feb. - 1 Mar. 1996, pp. 516–523.
121. J. Wilkinson and B. Devlin, "The material exchange format (mxf) and its application," *SMPTE journal*, vol. 111, no. 9, pp. 378–384, 2002.
122. L. Ying, S. Narayanan, and C. Kuo, "Content-based movie analysis and indexing based on audiovisual cues," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 8, pp. 1073–1085, Aug. 2004.
123. A. L. Yuille, "Energy functions for early vision and analog networks," *Biological Cybernetics*, vol. 61, pp. 115–123, 1989.
124. A. L. Yuille and H. H. Bülthoff, *Perception as Bayesian Inference*. Cambridge University Press, 1996, ch. Bayesian Decision Theory and Psychophysics, pp. 123–161.
125. J. M. Zacks, T. S. Braver, M. A. Sheridan, D. I. Donaldson, A. Z. Snyder, J. M. Ollinger, R. L. Buckner, and M. E. Raichle, "Human brain activity time-locked to perceptual event boundaries," *Nature Neuroscience*, vol. 4, no. 6, pp. 651–655, June 2001.

126. T. Zhang and C. C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 4, pp. 441–457, 2001.