To Appear in Proceedings of the IEEE, 2010

Audio-visual Information Fusion In Human Computer Interfaces and Intelligent Environments: A survey

Shankar T. Shivappa, Mohan M. Trivedi, and Bhaskar D. Rao

***Abstract*—Microphones and cameras have been extensively used to observe and detect human activity and to facilitate natural modes of interaction between humans and intelligent systems. Human brain processes the audio and video modalities extracting complementary and robust information from them. Intelligent systems with audio-visual sensors should be capable of achieving similar goals. The audio-visual information fusion strategy is a key component in designing such systems. In this paper we exclusively survey the fusion techniques used in various audio-visual information fusion tasks. The fusion strategy used tends to depend mainly on the model, probabilistic or otherwise, used in the particular task to process sensory information to obtain higher level semantic information. The models themselves are task oriented. In this paper we describe the fusion strategies and the corresponding models used in audio-visual tasks such as speech recognition, tracking, biometrics, affective state recognition and meeting scene analysis. We also review the challenges and existing solutions and also unresolved or partially resolved issues in these fields. Specifically, we discuss established and upcoming work in hierarchical fusion strategies and crossmodal learning techniques, identifying these as critical areas of research in the future development of intelligent systems.**

***Index Terms*—Multimodal systems, Information fusion, Audio-visual fusion, human activity analysis, Machine learning, Hidden Markov models, Dynamic Bayesian networks, Human activity modeling.**

I. Introduction

THE Turing test [1] suggests that a machine can be considered intelligent if a human judge involved in a natural conversation with a human and the machine, cannot distinguish between the two. The field of artificial intelligence, thus has its roots in making machines human-like, with the ability to perceive, analyze and respond to their surroundings in a way that is natural and seamless to humans. Since human perception is multimodal in nature, with speech and vision being the primary senses, significant research effort has been focussed on developing intelligent systems with audio and video interfaces[2].

The traditional interfaces such as keyboard, mouse and even close-talking microphones are considered too restrictive to facilitate natural interaction between humans and computers. Research efforts have been focussed on developing non-intrusive sensors such as cameras and far field microphones so that humans can communicate through natural means like conversational speech and gestures, without feeling encumbered by the presence of sensors. In other words, the computer has to fade into the background, allowing the users of the intelligent systems to conduct their activities in a natural manner. This necessitates the use of multimodal, especially audio-visual systems. Audio-visual systems are not restricted to human computer interfaces (HCI) alone. In several applications such as meeting archival and retrieval and human behavioral studies, audio-visual fusion can be applied as a post processing step. The techniques covered in this survey are also applicable in this context and not restricted to real-time interfaces.

Another significant advantage of using multimodal sensors is the robustness to environment and sensor noise that can be achieved through careful integration of information from different types of sensors. This is particularly true in cases where a particular human activity can be deduced from two or more different sensory cues, like for example, audio and lip movements in the case of human speech. Many other tasks like person tracking, head pose estimation, affective state analysis also exhibit significant overlap in the information conveyed over multiple modalities, especially audio and video.

Though different sensors might carry redundant information as suggested in the previous paragraph, these sensors are rarely equal, in the sense, they carry complementary information too, making it advantageous to use certain sensors over others for certain tasks. This is clearly demonstrated in the case of speech and gesture analysis for HCI applications, where the information carried through gestures complements the information presented through speech. Utilizing both these cues leads to a system that can understand the user more completely than using just one of the modalities.

There is yet another advantage of multimodal systems that has not been explored much in the existing literature. It is the ability to utilize the different modalities to learn cross-modal correspondences in an unsupervised manner. This has been studied in much detail in human cognitive studies[3][4]. This line of research holds promise in allowing us to exploit one of the biggest advantages of multimodal systems[5].

In this survey, we restrict our attention to systems that integrate multimodal sensory information to recognize human activities. Several recent surveys have been published in the area of human computer interfaces. Past surveys have looked into general information fusion schemes [6]. However, we are not aware of any survey that is wholly dedicated to sensory information fusion techniques used in multimodal

The authors are with the Department of Electrical and Computer Engineering, University of California - San Diego, La Jolla, CA 92093.

E-mail: sshivappa@ucsd.edu

To Appear in Proceedings of the IEEE, 2010

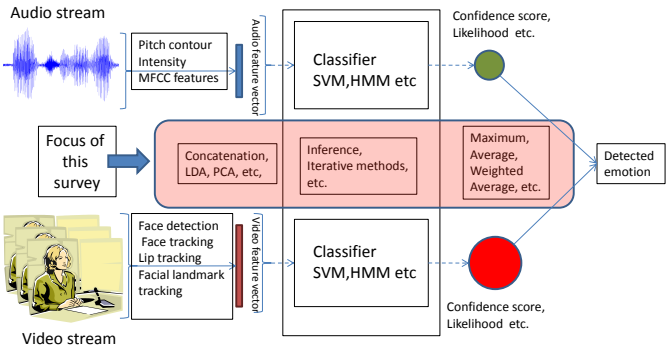


Fig. 1. An audio-visual emotion detection example illustrating the key issues in multimodal emotion detection and highlighting those discussed in this survey.

human activity analysis systems. The application domains of such systems are varied and include meeting analysis, robotics, biometrics, smart rooms etc.

Past surveys indicate the necessity for further research in the fusion schemes [7] [8]. Through this survey we hope to fill in the gap that exists in the systematic understanding of the challenges and existing solutions for fusion of sensory information across varying application domains. We critically compare the models and algorithms used to fuse information from multimodal, primarily audio-visual, sensors for human activity analysis. We also present some recent work from related fields like cognitive sciences that are relevant to addressing some of the challenges in current multimodal human activity analysis systems. In the present survey, we also restrict our attention to multimodal systems which use audio and visual sensors. However, the fusion techniques described are in general applicable to a variety of multimodal systems and hence wherever appropriate we refer to multimodal systems as opposed to audio-visual systems. In Figure 1, we use the example of an audio-visual emotion detection scheme to illustrate some key points of information fusion covered in our survey. There are other components of interest in these schemes such as sensor configuration, feature selection, modeling and training. It is almost impossible to totally separate these issues from one another and from the fusion strategies used. However, for the purpose of this survey, we try to focus on the fusion strategies alone. In the final sections of the survey, we focus our attention on critical future research directions in the areas of intelligent system design in general and human activity analysis in particular.

The rest of this paper is organized as follows. In section II, we discuss the theoretical and practical benefits of audio-visual fusion. We also outline a few concrete application areas where such fusion schemes can be very useful. In Section III, we propose a criterion for organizing the fusion strategies described so far in literature and we describe the fusion techniques in general, organized according to this criterion. We then compare some important fusion schemes used in practice in Section IV, organized according to their application domain. Finally in Section V, we highlight three key challenges that need to be addressed in the future and the ongoing efforts in that direction.

II. Benefits of Audio-Visual Fusion

Human activity and interaction is inherently multimodal. Vision and hearing are the primary senses used by humans to comprehend the complex world as well as to communicate with each other. Several psychological studies have outlined the fusion of audio and visual information by humans for performing particular tasks. A classic example is that of lip reading. Another example is that of audio source localization. These studies provide the basis for intelligent system researchers to incorporate either audio or visual or both the modalities in order to accomplish a particular task. It is necessary while designing such systems to evaluate the benefits and costs associated with using both audio and visual sensory modalities as opposed to using just one of them.

An example is the detailed analysis provided in Table II for the audio-visual speech recognition task. Such an analysis requires the collection of a standard dataset to evaluate the performance of the system. However, in practice, collecting such a dataset is extremely difficulty due to the varied nature of scenes and sensor configurations in which audio-visual fusion is applied. Consider the application of audio-visual speech recognition for a menu based system on an aircraft carrier deck with very high background noise levels. In such a situation, the performance depends on the configuration of the microphones as well as whether a head-mounted camera can be used to capture the lips of the speaker. In a meeting room however, the cameras and microphones are usually farther away from the users and though background noise is not much of an issue, far field sensors pose new challenges here. This is further complicated by the fact that the size and composition of the room affects the reverberation and visual clutter, posing different challenges to the audio and video processing. Inside a car cockpit, audio-visual speech recognition needs to address the issue of reverberation, back ground noise as well as overlapped speech from the other passengers. Thus there is a need to collect domain-specific audio-visual datasets as opposed to task specific datasets. For example, the CLEAR 2006 evaluation [9] and CLEAR 2007 evaluation [10] consider the domain of meeting scenes and then poses the question - what audio-visual fusion strategies can be employed in this context.

Cameras and microphones are ubiquitous and the current challenge is no longer the cost of deploying these sensors. Computational resources necessary to process the multiple data streams might be a limitation in some applications such as in mobile devices. However, the ongoing trends indicate that computational power will not be a bottleneck either. In order to be effective, an audio-visual system needs to use an effective fusion strategies. As outlined in the later sections of this survey, several audio-visual tasks, their corresponding suitable feature sets and fusion strategies have been explored by the research community. However, the selection of a suitable fusion strategy for a particular task at hand is non-trivial and requires domain expertise. This is a significant hurdle in the widespread deployment of audio-visual systems.

Future research needs to address this challenge by developing adaptive and context based fusion strategies. Online learning and automatic sensor calibration strategies will play a major role in the next generation of audio-visual systems.

A. Application domains

As discussed in the previous section, the selection of audio-visual fusion strategies is specific to the scene and sensor configuration. In this section we will briefly outline a few practical application domains.

The most extensively researched domain is that of meeting scenes. The challenge here is to use far-field cameras and microphones to analyze the human activity in a meeting scene, which typically has multiple subjects. A practical example of meeting analysis can be seen in [11]. Far-field sensors are necessitated for developing an unobtrusive system. Tasks such as person tracking, speech recognition, speech enhancement and person identification are performed using audio and visual cues.

Natural human computer interfaces are another domain where audio-visual fusion is critical. Here again, far-field sensors are used, however, the subject is usually co-operative and frequently adapts to the system.

Health smart homes and assisted living for people with disabilities is yet another area where audio-visual systems are needed [12]. This includes passive surveillance of the scene for detecting certain events of interest such as an individual losing consciousness/mobility as well as active interaction with the subjects.

Intelligent vehicles have advanced significantly and include several driver assistance technologies [13][14]. Such driver assistance systems and the interaction with the car's infotainment system could benefit significantly by the use of both audio and visual cues[15]. Speech recognition, person identification, affect analysis are tasks of interest in this context.

Several psychoanalytical studies involve the segmentation and labeling of audio-visual recording of subjects. Using audio-visual fusion framework to develop segmentation algorithms has a great potential in making this process more efficient and affordable.

In figure 2, we present audio-visual testbeds involving meeting scenes, natural HCI and intelligent vehicles. Though audio-visual fusion is not commonly employed in the real-world applications at present, there is a lot of potential that needs to be explored and these testbeds are a first step in that direction.

III. Audio-visual Information Fusion Schemes

The varied application domains of multimodal human activity analysis systems have always presented a challenge to the systematic understanding of their information fusion models and algorithms. The traditional approach to information fusion schemes classifies them based on early, late and intermediate fusion strategies and describes their associated merits. However most multimodal systems are built to exploit one or more of the advantages as described in Section I. Thus a classification of the fusion strategies based on their "intent" would provide a new angle to look at these schemes. Also, the various models used, probabilistic and otherwise also need to be examined for their merits in the fusion schemes. In this survey, we organize the existing research in multimodal information fusion schemes based on these criteria.

Humans are the ultimate intelligent systems equipped with multimodal sensors and the capability to seamlessly process, analyze, learn and respond to multimodal cues. Humans beings seem to learn the cross-modal correspondences early on and use that along with other techniques to combine the multimodal information at various levels of abstraction. This seems to be the ideal approach to sensory information fusion as exemplified by the success of hierarchical modeling schemes. However significant progress is necessary before computers can begin to process multimodal information at the level of humans. The models and algorithms used in intelligent systems need not be motivated by human information processing alone. However, human cognition can provide valuable insight into the what and how of intelligent systems.

It is extremely challenging to organize the extensive literature in the field of multimodal systems in a precise and useful manner, in order to be able to extract meaningful information from it. In this survey, we aspire to systematically classify and study the various multimodal information fusion schemes reported in the literature of human activity analysis so far. In this section we organize the multimodal fusion schemes based on their primary "intent", that is, the primary reason among those specified in Section I that the researchers have tried to address in designing the system. For example, an audio-visual speech recognition system's intent would be to achieve robustness to environmental and sensor noise. Further, we explore the traditional early/intermediate/late fusion strategies and the different modeling techniques used under each of these main categories. Thus the systems are classified as those that use multimodal sensors primarily for

- **Achieving robustness to environmental and sensor noise.**
- **Facilitating natural human computer interaction.**
- **Exploiting complementary information across modalities.**

Achieving robustness to environmental and sensor noise is the traditional motivation for audio-visual information fusion. Thus this category includes the major part of the multimodal fusion strategies studied so far. The most widely accepted notion of sensory information fusion applies to these systems. Those tasks which involve redundant cues in multiple modalities due to the nature of the human activity, fall under this group. Audio-visual speech recognition is the classic example of such a task. It is also one of the earliest areas to generate considerable research interest in multimodal information fusion techniques. We also cover other areas including audio-visual person tracking, affect analysis, person identification etc. However, the organization will be based on the fusion strategies rather than the particular application domain. In earlier literature[16][7], fusion strategies have been classified as follows -

- **Signal enhancement and sensor level fusion strategies.**

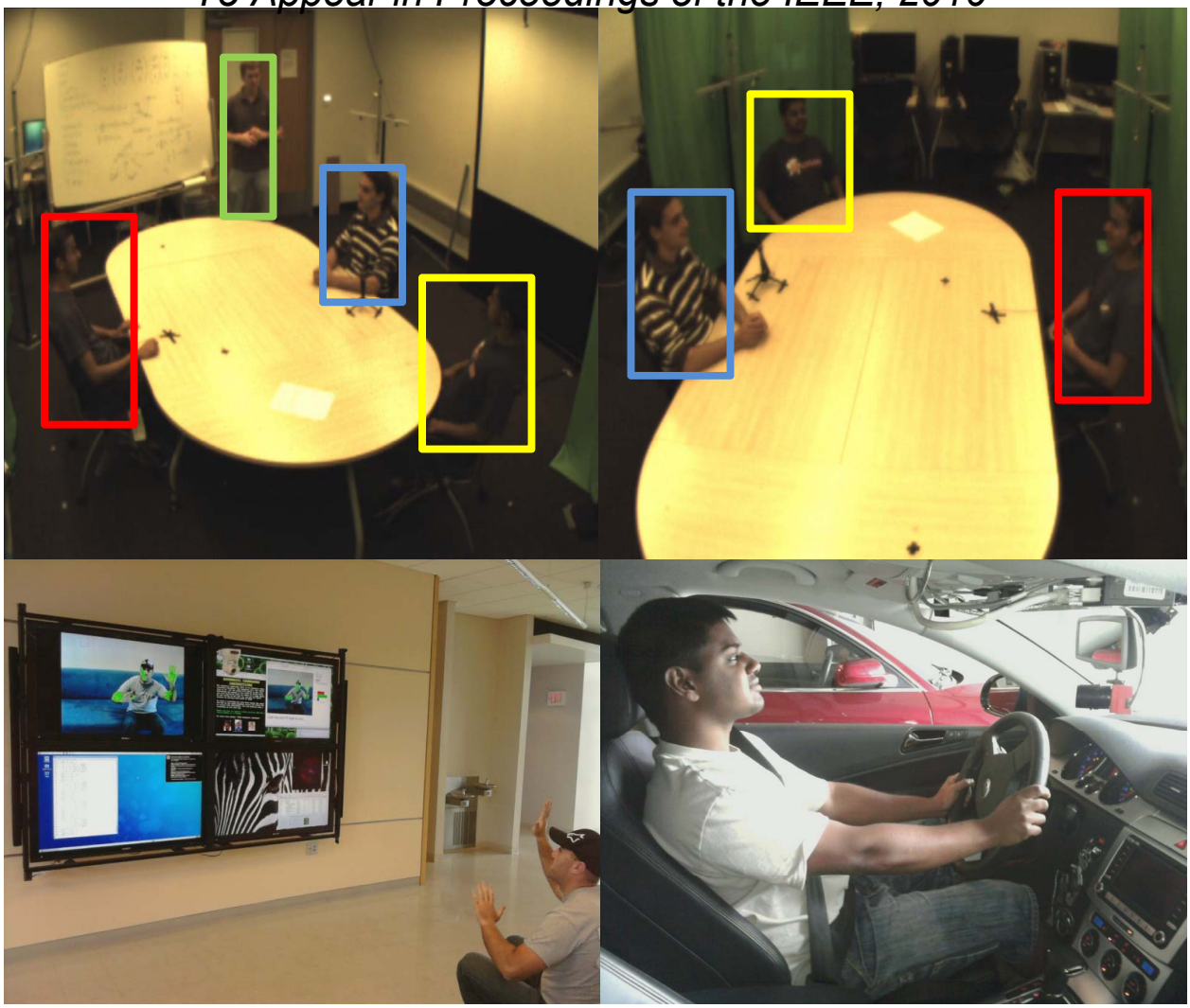


Fig. 2. Audio-visual testbeds involving a meeting scene, natural HCI and an intelligent vehicle. These are some of the examples where the benefits of audio-visual fusion are being demonstrated in real-world situations.

- **Feature level fusion strategies.**
- **Classifier level fusion strategies.**
- **Decision level fusion strategies.**
- **Semantic level fusion strategies.**

In the following sections we will explore systems that utilize these strategies and discuss their relative merits and de-merits.

A. Signal enhancement and sensor level fusion strategies

This includes signal enhancement techniques such as beam-forming using microphone arrays. It is conceivable that video information could be useful in the beamforming process as in [17][18][19]. Also, camera networks could benefit from the source localization and pan-tilt-zoom cameras might be able to capture better images of the scene. However such schemes are rarely described in isolation and we present more examples in later sections (V-A) that deal with hierarchical approaches.

B. Feature level fusion strategies

Cognitive scientists refer to this as the early fusion strategy. This is also referred to as the data to decision fusion scheme in literature[20][21]. Some tasks such as automatic speech recognition, person tracking, affect analysis etc produce cues in multiple modalities in a temporally correlated manner. Note that an up-sampling or a down-sampling stage is sometimes necessary in order to align the streams to each other. A representative example is the case of audio signals and lip movements carrying the information about the spoken word in the audio and visual modality respectively. In these cases, an early fusion strategy is feasible. In this case, one concatenates the feature vectors from the multiple modalities to obtain a combined feature vector which is then used for the classification task. Figure 4 is a schematic representation of typical feature fusion schemes.

This early fusion has the advantage that it can provide better discriminatory ability for the classifier by exploiting the covariations between the audio and video features[20]. However, the larger dimensionality of the combined feature vector presents challenges for the classifier design. In order to overcome this, standard dimensionality reduction techniques such as DCT, PCA, LDA and QDA are applied. LDA and QDA

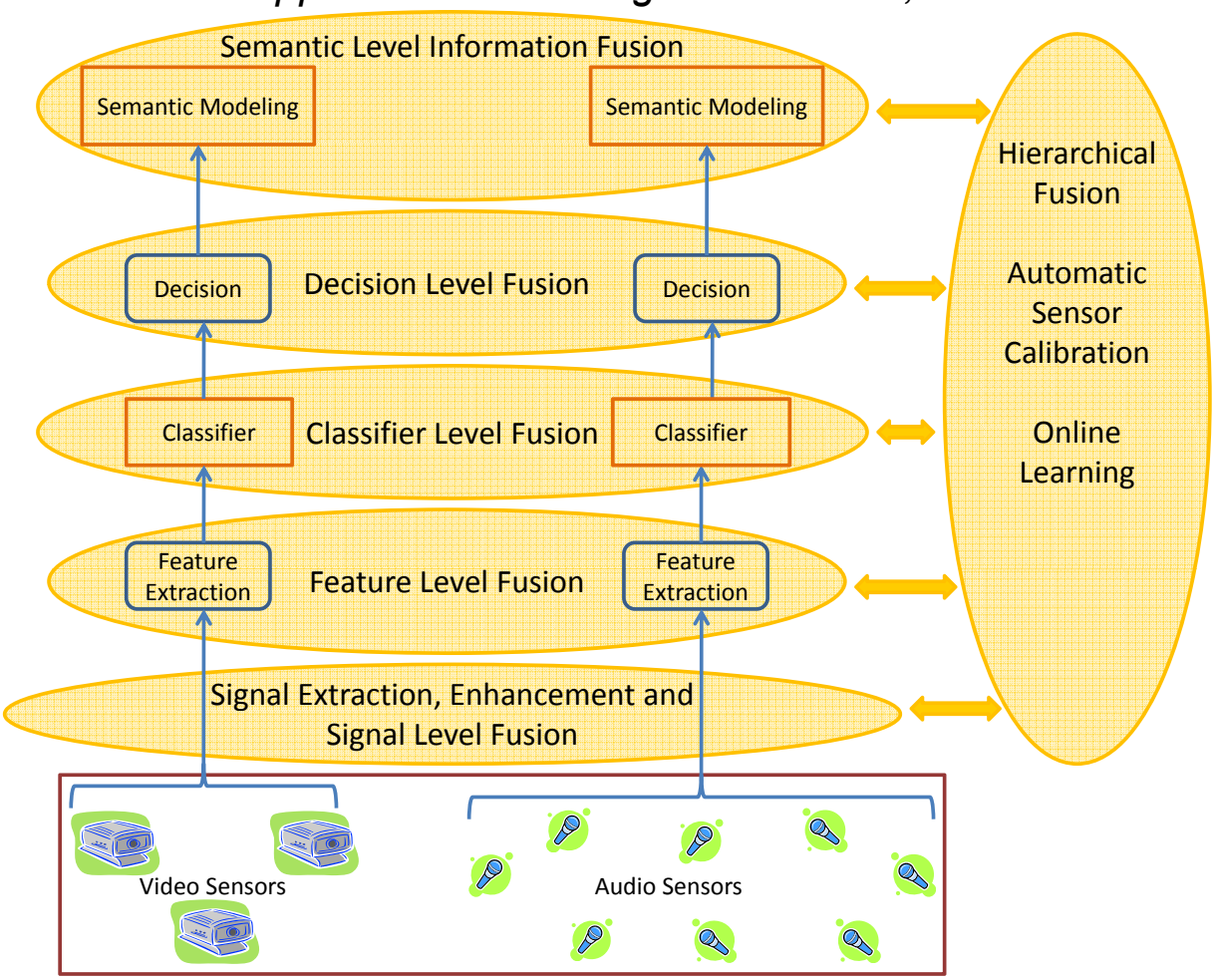


Fig. 3. Information fusion at various levels of signal abstraction is depicted here.

based systems are know to out-perform PCA based systems in classification tasks. However, in the presence of limited training data, PCA is more stable than LDA[22]. The optimal dimensionality reduction technique also depends on the nature of the classifier used. Theoretically, kernel based classifiers like SVMs do not require an explicit dimensionality reduction step. However, most multimodal systems adopting an early fusion strategy are based on HMM based classifiers and do benefit from dimensionality reduction.

As an example, [23] presents an elaborate scheme for early fusion of audio-visual information for speech recognition which includes both early and late fusion. The early fusion consists of the DCT and multiple PCA steps to reduce the dimensionality of the audio-visual feature vector. Early fusion strategy with a HMM based classifier is also explored in [24] for the purpose of analyzing group actions in meetings. 39 features including 18 audio features and 21 visual features are concatenated and used to recognize group actions in meeting recordings. This early fusion scheme is second in performance only to an intermediate fusion strategy using asynchronous HMMs (10% vs 9.2% error rates), revealing that the simple early fusion strategy is quite effective if used in the right task. Another example of early fusion for audio-visual tracking can

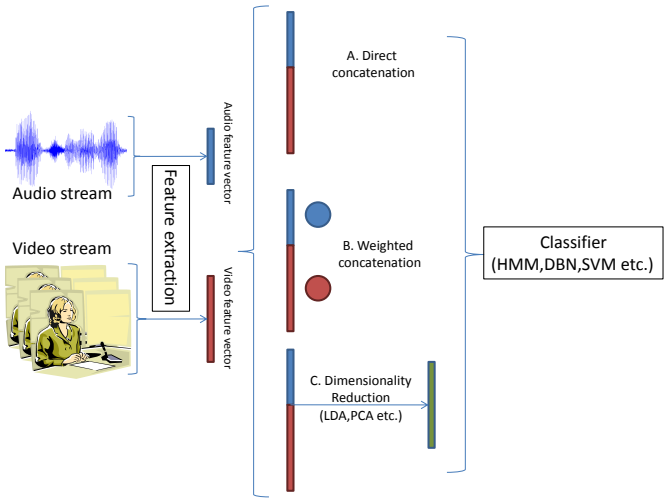


Fig. 4. Signal and information flow in feature level fusion strategy

be seen in [25]. Here the microphone arrays and cameras are treated as generalized directional sensors and treated equivalently. [26] proposes an iterated extended Kalman filter (IEKF) for audio-visual source tracking by concatenating audio and

visual features.

The early fusion technique has the advantage of being the simplest to implement and is suitable for those applications which require very fast processing of cues. However, it cannot be applied to most tasks where strictly temporally synchronized cues are not present. Also, the feature concatenation performs poorly when the reliability of the different modalities during the training phase differ from the actual operation phase.

C. Classifier level fusion strategies

Cognitive scientists refer to this as the intermediate fusion strategy. This is typically encountered in cases where HMMs (and their hierarchical counterparts) and Dynamic Bayesian networks are used to model individual streams. In such cases, the information can be fused within the classifier, but after processing the feature vectors separately. Thus a composite classifier is generated to process the individual data streams. The intermediate fusion strategy is an attempt to avoid the limitations of both early and late fusion strategies . Unlike early fusion, fusion at the classifier level does allow the weighted combination of different modalities based on their reliability[27]. These weighted combinations however are taken on each frame, allowing for a much finer combination of cues than in late fusion. Such fusion schemes are widely used in audio-visual speech recognition systems. Figure 5 is a schematic representation of typical intermediate fusion schemes.

Asynchrony between the different streams can be modeled to some extent. This is critical in cases such as audio-visual speech recognition where the audio and video asynchrony is of the order of 100ms whereas the frame duration is typically 25ms[28][16]. Different degrees of asynchrony are allowed at the cost of complexity and speed. The multistream HMM[29][30] assumes perfect synchrony between the different streams. On the other extreme is the model that allows complete asynchrony between the streams. This is however infeasible due to the exponential increase in the number of state combinations possible due to the asynchrony. An intermediate solution is given by the product HMM [31] or the coupled HMM [32]. In case of audio-visual speech recognition, this corresponds to imposing phone synchrony as opposed to the frame synchrony of the multistream HMM.

The coupled hidden Markov model and the multistream hidden Markov model have been used to improve the performance of audio-visual speech recognition [29], [32], [33]. These schemes have also been applied in other areas of research such as biometrics[34], audio-visual head pose estimation using the particle filter framework[35], audio-visual person tracking[36][37][38][39], audio-visual aggression detection[40]. However, in the real-world situations, the reliability of the different streams varies with time. For example, the video channel in audio visual speech recognition might be completely unreliable if the speaker covers the mouth with the hand or turns away from the camera[41]. In this case, it is useful to be able to estimate the reliability of each channel continuously and weight them accordingly. Stream weight

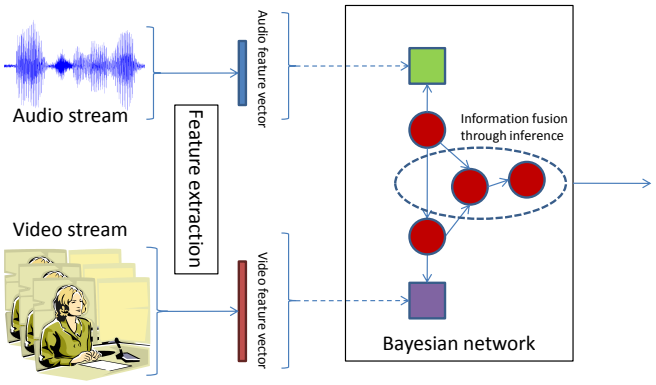


Fig. 5. Signal and information flow in classifier level fusion strategy

estimation and its adaptive counterparts have been presented in literature[16][42]. The iterative decoding algorithm [41] solves this problem by using techniques burrowed from turbo codes [43]. The iterative decoding algorithm has been applied to the problem of audio visual speech recognition[33] on the GRID audio-visual speech corpus[44] and to the problem of person tracking using the audio-visual cues in [37].

D. Decision level fusion strategies

Late or decision level fusion involves the combination of probability scores or likelihood values obtained from separate uni-modal classifiers to come up with a combined decision. In cases where strictly temporally synchronized cues are absent, late integration is still feasible. Typically late fusion involves using independent classifiers, one for each modality and combining the likelihood scores based on some reliability based weighting scheme. The training and decoding these uni-modal models scales linearly in the number of streams which makes these schemes particularly attractive. The reliability of the streams is typically used by exponentially weighting the probability scores from individual streams before taking their product. Such a combination scheme with appropriate weighting scheme has been used for audio-visual speech recognition[29]. However, in case of audio-visual speech recognition, the late fusion strategy has been shown to be inferior to the intermediate fusion strategy discussed in the previous section[29]. Figure 6 is a schematic representation of typical decision fusion schemes.

The weighting scheme used in late fusion draws upon the work in combination theory to estimate the best weighting factors based on the training data. This is however a limitation when there is a mismatch between the training database and the actual operation. As with the intermediate fusion strategy, decision fusion allows for separate weighting of the different streams based on the reliability. However the fusion is not at the level of frames but at a higher levels. For example, in the audio-visual speech recognition context, the decision level fusion could take place at the utterance level. Decision level fusion allows maximum flexibility in the choice of individual classifiers. [45] explores the use of decision level fusion for audio-visual person identification. The lack of state correspondences in the text independent person ID task imposes the late

To Appear in Proceedings of the IEEE, 2010

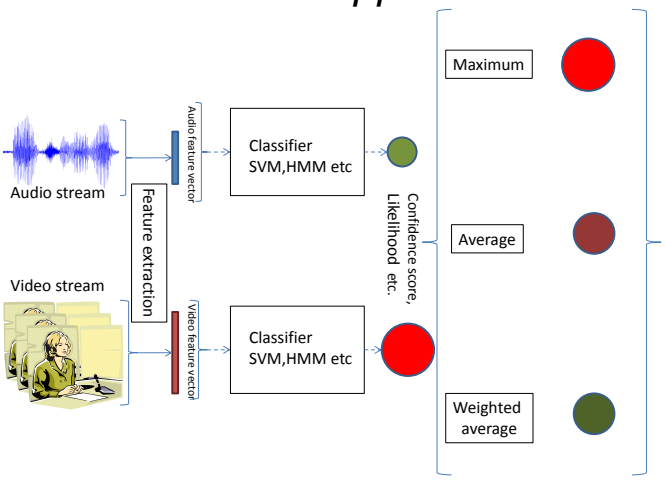


Fig. 6. Signal and information flow in decision level fusion strategy

fusion strategy in this case. The authors also acknowledge the importance of optimal weighting in the decision fusion. [46] is another audio-visual person identification system based on decision level fusion. [47] describes an audio-visual affect recognition which uses decision level fusion to combine facial expressions and prosodic cues for affective state recognition.

E. Hybrid fusion strategies

A combination of the above mentioned fusion strategies is also reported in literature. In [16] a combination of feature level fusion with decision level fusion is used in the context of an audio-visual speech recognition task. The audio and visual feature are combined early on through a discriminatory feature selection process and the discriminatory features are used again as one of the streams in a multi-stream based decision fusion technique. There is no theoretical basis for such a scheme, however in practice, it is shown to improve the recognition accuracy. Canonical correlation analysis (CCA) is a statistical approach that combines linear dimensionality reduction and fusion by computing linear projections that are maximally correlated. It is a combination of early and late fusion strategies. [48] applied CCA to a open-set speaker identification problem. More recently, a spectral diffusion framework has been proposed to provide a uniform embedding of data for multisensory fusion [49].

F. Semantic level fusion strategies

It is conceivable that higher level information can be merged after the semantic interpretation of the sensory information. This is beyond the scope of our survey because usually such fusion schemes will involve other modalities like text, webpages and other such sources of information that are amenable to semantic interpretation.

G. Facilitating natural human computer interaction

This is yet another category of audio-visual fusion strategies that we will review briefly. In this category we include those systems which use multimodal sensors with the primary intention of facilitating natural modes of interaction for the users. As such, these systems do not focus on the actual fusion of multimodal sensory cues, with emphasis being more on the interaction part rather than sensing part. Nevertheless the systems use multimodal sensors and incorporate some aspects of fusion. Human-robot interaction [50][36] and avatar based HCI[51][52][53] are the most popular applications in this category. As an example, [54] and [55] describe the anthropomorphic robot 'Kismet' which uses vision based processing to perceive user intentions and a microphone to enable vocal interactions in a turn-taking conversation. The audio and visual information streams are not merged explicitly but together enable the system to perform the task of carrying out a sociable conversation.

H. Exploiting complementary information across modalities

The third main category of audio-visual fusion techniques are reviewed in this section. Context aware intelligent spaces[56][57] are good example of systems that can exploit complementary information across modalities by extracting information from different sensors based on the context. Different sensors provide information about different aspects of the overall task/goal of the system. Another example of a system using such a fusion strategy is the gesture driven speech recognition interface. Here, the gestures compliment the information provided through speech. In both these cases, the information carried in different modalities are generally disjoint and are integrated at higher levels of abstraction to achieve a comprehensive task. Systems modeling group interactions and meetings[58][59] typically fall under this category. Human Robot interaction is another area in which speech and gestures are usually combined with the gestures complementing the information obtained from speech commands. [60] describes a system that integrates spoken natural language and natural gesture for command and control of a semi-autonomous mobile robot.[61] is another example of using vision and speech where a finite state machine structure is used to enable the robot to learn a map of the area using human assistance. The commands are speech based but human detection and hand gesture recognition are used to resolve some ambiguities in the speech commands and to select and localize objects referred to by speech.

IV. Comparison of Fusion Strategies on a Task Specific Basis

In the preceding sections, fusion strategies are described in general for the various audio-visual tasks. An important observation is that the modeling technique and task specific details determine the fusion algorithms used. The modeling techniques are themselves task dependent. For example, researchers from the audio-visual tracking community prefer to use Bayesian networks with particle filter based inference [62][63][64][65][66]. In the following sections, we review and compare the key contributions in fusion techniques on a task specific basis.

TABLE I
RELATIVE MERITS AND DEMERITS OF TRADITIONAL CATEGORIES IN AUDIO-VISUAL INFORMATION FUSION STRATEGIES.

Fusion scheme	Noiseless case	Noisy case	Modeling restriction
Feature level	Can use the co-variations in data streams	Cannot separate out the noisy features	Applicable to most of the generic classifiers
Classifier level	Cannot use the co-variations	Can separate the influence of the noisy modality and fuse the appropriate information at each frame	Applicable to specific models like HMMs and DBNs
Decision level	Cannot use the co-variations	Can separate the influence of the noisy modality but the fusion is at a coarser level	Applicable to most generic classifiers

A. Fusion strategies in audio-visual speech recognition tasks

Audio-visual speech recognition is one of the first tasks to highlight the fusion of audio and visual information to achieve robustness to background noise and to improve the recognition accuracy by including additional visual information to the audio based speech recognition. Speech is bimodal in nature [67][68][2]. Audio based automatic speech recognition is extremely sensitive to background noise. Incorporating visual information provides robustness to background audio noise. The fusion schemes have evolved with the modeling techniques and are mostly based on appropriate weighting of the audio and video streams based on the level of background noise. At very high audio SNR, audio-only speech recognition performs as well as the audio-visual recognizer. At very low SNR, the performance is worse and closer to that of the video-only speech recognizer. In between these two extremes, the aim of the fusion schemes has been to achieve a graceful degradation in performance as the SNR decreases. Another challenge in audio-visual speech recognition is the lack of one-to-one correspondence between the phonemes and visemes. Yet another important aspect is the difficulty in extracting clean audio and video features from the speaker. Speech is a natural mode of interaction for humans and the use of lapel microphones and cameras with frontal view of the speakers makes the speech recognition interface restrictive. Microphone arrays and visual face tracking using multiple cameras have been used to undo this restriction to some extent. Also, the recognition accuracy for conversational speech is quite low and thus restricts the usage of automatic speech recognition in meeting scenes. Further research in audio-visual strategies is necessary to tackle these limitations. Some key contributions have been summarized in Table II.

Initial work in "speech reading" [69] was based on binary images of the mouth region and involved a sequential recognition strategy. The audio recognizer was first used to come up with a candidate list of words and the visual features were used on this list to arrive at the final decision. This system was later improved in [70] by utilizing rule based heuristics to combine the audio and visual recognition results. The experiments were restricted to a limited vocabulary, speaker dependant speech recognition task. In [71], the authors explored a different strategy for fusion of information by using a neural network to map the visual features to the acoustic spectrum domain. The fusion is achieved by weighted averaging of the actual acoustic spectrum and the spectrum obtained from the visual features. Multi state Time delay neural networks (MS-TDNN) were used in [72] and reported 50% reduction in error rates in some cases with acoustic noise, on a speaker dependent task. This work extended the audio visual fusion scheme to connected letter recognition using the dynamic time warping(DTW) framework. The combination of the audio visual information at different levels of the neural network is also considered in [72] and the relevant weights are learned by back-propagation on the training set. [73] introduces an HMM based recognizer for audio-visual speech recognition. [74] investigates an HMM based speech recognizer with feature concatenation strategy on speaker independent connected digits task. The system performed as expected, with little change in the high SNR case. As the SNR decreased and the performance of the audio-only recognizer worsened, the visual information provided a gain of about 10dB in SNR. [75] describes the efforts towards a large vocabulary continuous speech recognition using audio-visual features and efforts to develop a database to facilitate the research in the same. These various fusion schemes are organized and are compared in [20] which also introduces a new fusion approach by mapping the acoustic and visual streams to the motor space. The results suggest that both early and late fusion schemes have similar performance and the performance of the motor recoding and dominant mode recoding schemes is worse.

Hidden Markov models(HMM) with Gaussian mixture model(GMM) observation densities are commonly used in more recent speech recognition systems for modeling and recognizing speech[76]. Several audio and video features have been proposed in literature but feature selection is outside the scope of this survey. [23] is the outcome of the 2000 Johns Hopkins University workshop on audio-visual speech recognition. The following models were evaluated and the results summarized in Table II.

- Early fusion technique by concatenating the audio and visual features - AV-Concat
- Early fusion technique with Hierarchical LDA for dimensionality reduction - AV-HiLDA
- Multistream HMM with jointly trained unimodal models[77] - AV-MS-1
- Multistream HMM with individually trained unimodal models - AV-MS-2
- Multistream HMM with jointly trained unimodal models

and utterance specific stream weights - AV-MS-UTTER
- Multistream Product HMM which allows for state asynchrony between modalities[29] - AV-PROD)
- Late integration with discriminatory model combination[78] - AV-DMC

The results show that on the speaker-independent audio-visual large-vocabulary continuous speech recognition task, Av-HiLDA, AV-MS-UTTER and AV-PROD provided consistent improvement for both clean and noisy audio conditions considered. The multistream HMM based fusion schemes performed slightly better than the feature fusion schemes. The choice of the exponents for the different streams in multistream HMM is not completely resolved. Only the AV-MS-UTTER estimates the exponents based on the degree of voicing in the utterance. The rest of the schemes rely on estimating the optimal weights based on a training set. But in practice, the noise levels are not known apriori.

[79][80][81] use a probabilistic descent algorithm on word classification error on held-out data for stream exponent maximization. These and related techniques are not suitable for practical systems. The quality of the audio and video streams varies with time and it is not optimal to estimate the stream reliability exponents based on a predetermined training set. [16] addresses the issue of stream reliability estimation from the observations without the use of annotated training data. Given the observations of individual streams, the class conditional likelihoods of their N-best most likely generative classes are used to estimate the N-best likelihood difference and the N-best likelihood dispersion. These measures are well correlated to the WER and are used to estimate the stream weighting exponents. Important stream weight selection criteria have been summarized in Table IV.

Another issue that needs to be incorporated in the modeling scheme is the relative asynchrony between the audio and visual cues. [82] compares the performance of the coupled HMM and factorial HMM on a large vocabulary continuous audio-visual speech recognition task. The coupled HMM performs better than the factorial HMM. Both the models are treated as specific instances of dynamic Bayesian networks. DBNs have also been used to model auxiliary information in speech such as articulator positions [83][84][85]. More recently discriminatory techniques such as boosting and SVM have been employed in hybrid architectures to improve the recognition accuracy[86][87]. In practice however, audio-visual speech recognition remains a challenging problem due to the time varying nature of the audio and visual streams[23]. This could involve the conditions like non-stationary background noise in the audio domain and occlusions and changes in position and orientation of the human head in the visual domain. [33] explores the use of the iterative decoding scheme to combine the audio-visual cues. In the iterative decoding scheme, the reliability of the individual modality is implicitly modeled by the distribution of the extrinsic information passed back and forth between the two modalities.

The schemes described so far assume the availability of close talking microphones and cameras with frontal view of the speaker. This is a severe restriction in many cases such as meeting scenes where the participants' position and orientation varies with time. Multiple cameras and microphone arrays have also been studied to localize and extract robust features for speech recognition[88][89][19]. These systems use the audio-visual speaker localization to enhance the speech signal using beamforming techniques. This line of research has its limitations based on the room acoustics [90], environmental noise conditions and the overlapping speech during meetings. Thus a combination of the speech enhancement techniques for far-field microphone arrays, augmented with visual information in the form of lips and contextual information promises to yield fruitful results.

B. Fusion strategies in audio-visual person localization and tracking

Person tracking has been a computer vision problem that received considerable attention[93]. Audio source localization is also a well researched field [90][94]. Localizing and tracking individuals using audio-visual information has recently received much attention. Some key contributions have been summarized in Table V.

Camera epipolar constraint and microphone array geometry based schemes have been reported [95] [96] [97].[98] presents a skin tone based algorithm for omni directional cameras. The HMM based tracker has been used in [99] in conjunction with the iterative decoding scheme. [100] describes a audio localization system for camera pointing which uses audio-visual correspondences to calibrate a microphone network for source localization relative to the camera. Audio-visual synchrony and correlation have been exploited to locate speakers in [65][101][102]. Bayesian networks with the particle filtering based inference technique have been widely used in audio-visual tracking [103] [104] [105] [66] [64] [63] [38] [26] [62] [39]. Approximate inference in the dynamic Bayesian network framework, necessitated by the complexity and non-Gaussianity of the joint models, is performed by the use of particle filters [62],[63]. An HMM based iterative decoding scheme is presented in [99]. In Table V, we see that the scene and sensor configurations are varied and the results are not presented on any standard dataset. Hence it is not possible to compare the different frameworks. The CLEAR 2006 evaluation [9] and CLEAR 2007 evaluation [10] address this issue by providing a standard dataset to evaluate person tracking frameworks. In [99], the authors compare the iterative decoding framework with the particle filter framework. The results are presented in Figure 7. Note that the particle filter framework which tracks in the 3D co-ordinates is sensitive to accurate sensor calibration as opposed to the iterative decoding scheme which tracks in the local sensor coordinates. Calibration of the sensors with respect to the 3D world co-ordinates is an important issue in person tracking systems. We do not discuss the different calibration schemes in detail in this review. However, automatic calibration of sensors is a major advantage of audio-visual systems and this aspect needs to be explored further.

We observe that earlier work on audio-visual tracking involved the development of a framework for data association and tracking based on audio and video cues. More recent work

TABLE II

THE ACCURACY OF DIFFERENT FUSION SCHEMES FOR AUDIO-VISUAL SPEECH RECOGNITION HAVE BEEN COMPARED ON A LVCSR TASK : SUMMARY OF RESULTS FROM [23]

Model	Clean	Noisy	Model	Clean	Noisy
Audio-only	85.56%	49.90%	AV-MS-1	85.38%	63.39%
AV-Concat	84.00%	60.00%	AV-MS-2	85.08%	62.62%
AV-HiLDA	86.16%	63.01%	AV-MS-PROD	85.81%	64.79%
AV-DMC	86.35%	—	AV-MS-UTTER	86.53%	64.73%

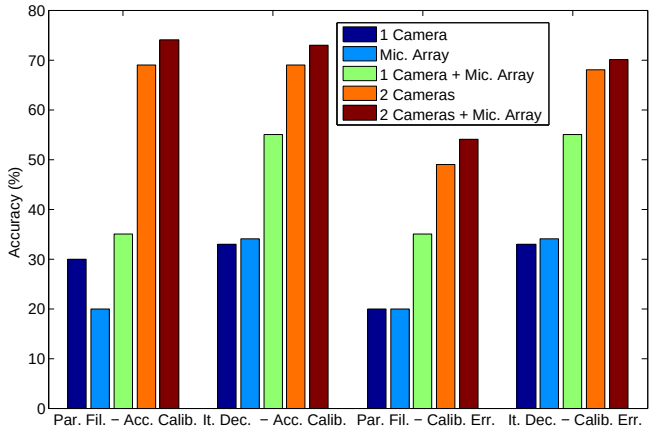


Fig. 7. Results from [99] - percentage of occlusions that are correctly resolved by the iterative decoding framework in comparison with the Particle filtering based tracker. Note that the performance of the two schemes is very similar when accurate calibration is available whereas the performance of the particle filter based tracker degrades in the presence of calibration errors.

includes the use of hierarchical fusion schemes for augmenting the tracks with additional information. In [39][17] person ID is combined with the tracks to improve the accuracy of tracking. Audio-visual tracking is a fundamental task in any human activity analysis system. However, fusing audio-visual cues for person tracking is extremely challenging and will likely include many different kinds of cues such a speaker identification and face recognition (Section V-A).

C. Fusion strategies in audio-visual affective state recognition and emotion recognition

Early research in audio-visual affective state recognition [106] was based on rule based classification of audio visual cues into one of the six categories: happiness, sadness, fear, anger, surprise and dislike. Pitch, intensity and pitch contours were utilized as acoustic features, whereas facial features like eyebrow position, opening of eyes and mouth and frowning are measured and classified based on a set of rules. [107] used similar features but classified visual features using a nearest neighbor classifier and the audio features using an HMM. If the audio and video results did not concur, the decision of the dominant mode prevailed. The dominant mode for each emotion was selected based on the training data. [108] used an ANN classifier for visual classification and HMM for audio emotion recognition. The resulting scores were averaged before making the final decision. The above research considered person dependent models. In [109], the authors investigated both person-dependent and person-independent models. The person-dependent model achieved a recognition accuracy of 79% whereas the person-independent model achieved 56% accuracy.

Audio visual feature extraction for emotion detection is a very challenging field in itself but is outside the scope of this survey.

More recent work includes the use of discriminative classifiers like SVMs[110][111] and boosting and more expressive generative classifiers like Bayesian networks [112] to fuse audio-visual information. Table VI lists the details of some key fusion schemes reported in literature. There are a lot of open challenges in affective state analysis. One specific issue that is very evident in emotion detection is the temporal offset between the expression of the audio and visual emotional cues. This temporal mismatch is much more pronounced in the features related to emotion detection than other audio-visual tasks. Another important challenge is the lack of suitable large emotion databases in American English. Moreover at its current stage, most of the systems are designed to work with close talking microphones and cameras with good frontal view. This places severe constraints on the users of such systems. Techniques need to be developed to use distributed array of cameras and microphones and detect the emotional state of subjects involved in mostly unconstrained activities.

D. Fusion strategies in audio-visual person identification

Biometrics is another active area of research for audio-visual fusion. Though biometrics several fingerprints, palm prints, hand and finger geometry, hand veins, iris and retinal scans, infrared thermograms, DNA, ears, faces, gait, voice and signature, the audio and video modalities facilitate unobtrusive, user friendly, lower-costing-sensor based person identification [117]. Moreover, multi modality is not only used to extract complimentary information to boost recognition accuracy, but is also helpful in making the person identification systems fool-proof against imposters attacks [118][119]. Audio-visual fusion has been used in both open-set[34] and closed-set[120] person identification tasks. [121] presents a good overview of the various information fusion schemes used in audio-visual person identification. The pre-classifier level, classifier level and post-classifier fusion examples are presented along with an evaluation of fixed and adaptive fusion techniques. The authors also propose a piecewise linear classifier and modified Bayesian post classifier that adapt the decision boundaries based on the environmental noise level. [122] presents another evaluation of fusion schemes but in the context of more general

TABLE III

SUMMARY OF FUSION SCHEMES IN AUDIO-VISUAL SPEECH RECOGNITION A - ISOLATED WORD/LETTER/DIGIT. B - CONNECTED LETTERS/WORDS/DIGITS. C - CONTINUOUS SPEECH. D - STATIC VOWEL. P - SPEAKER DEPENDENT. Q - SPEAKER INDEPENDENT. X - SMALL VOCABULARY. Y - LARGE VOCABULARY.

Fusion strategy for audio-visual speech recognition	Model	Domain	Performance - Audio - Visual - Audio-visual	Publication and Year
Sequential, starting with audio	Template matching	A,P,X	-	Petajan [69] 1984
Rule based	Vector quantization and dynamic time warping	A,P,X	63% 72% 86%	Petajan et. al. [70] 1988
Mapping visual features to acoustic spectrum and weighted averaging	Spectrum based classification	D,P,X	60% — 81%	Yuhas et. al. [71] 1989
Product rule	Time delay neural network (TDNN)	A,Q,X	43% 51% 75%	Stork et. al. [91] 1992
Weighted combination of beliefs in the NN	Multi state Time delay neural network (MS-TDNN)	B,P,X	47% 32% 76%	Bregler et. al. [72] 1993
Late integration with weighted probability scores	Hidden Markov Model (HMM)	A,P,X	70% 76% 96%	Adjoudani and Benoit[73] 1996
Late integration by weighted log-likelihood combination	Multistream HMM	B,P,X	19% 36% 86%	Potamianos and Graf[79] 1998
Classifier level fusion based on probabilistic inference	Dynamic Bayesian network (DBN)	A,Q,X	90% — 92%	Stephenson et. al. [84] 2000
Late integration by optimally weighted log-likelihood combination	Multistream HMM	B,Q,X	62% — 91%	Nakamura et. al. [80] 2000
see Table II		C,Q,Y		Neti et. al. [23] 2001
Weighted HMM-state likelihood combination	Coupled HMM	C,Q,Y	50% 67% 77%	Nefian et. al. [82] 2002
Hybrid fusion : feature and decision level fusion	Multistream HMM	C,Q,Y	60% —- 78%	Potamianos et. al. [16] 2003
AdaBoost	Boosted HMM classifier	B,P,X	66% 36% 73%	Yin et. al. [86] 2003
Classifier level fusion based on probabilistic inference	Dynamic Bayesian Networks	C,Q,X	76% 53% 88%	Gowdy et. al. [85] 2004
Late integration with weighted probability scores	Hybrid SVM-HMM	A,Q,X	80% 80% 91%	Gurban and Thiran[87] 2005
Iterative decoding	Multistream HMM	A,P,X	60% 72% 75%	Shivappa et. al. [33] 2008

multimodal biometric systems. Table VII lists the details of some key fusion schemes reported in literature. It is easily seen from the table that the weighted combination of classifier scores is a very popular fusion strategy. Person identification task is a very structured classification task and hence the polarization towards the particular fusion strategy. However the selection of weights is an open issue and researchers have addressed it using different techniques. Predominant are the ones based on classification error and the reliability of the modality which makes intuitive sense. However, an existing challenge is to incorporate person identification into a setup with distant microphones and cameras. Some groups have begun to address this issue [39][17]. Further research is necessary to bring out the strengths of person identification systems not only in biometrics applications but also in other human activity analysis. Person identification is an inherently multimodal task and will form the basis for future hierarchical information fusion schemes (Section V-A).

E. Fusion strategies in audio-visual meeting scene analysis

Audio-visual analysis of human activity in meeting rooms for meeting scene understanding, segmentation, archival and retrieval has received a lot of attention in the recent past. Systematic comparison of the different fusion approaches in meeting scene analysis is extremely challenging due to varying scenarios considered by different groups. Moreover, many of the systems described above (ASR, biometrics, tracking, emotion detection etc.) are used as subsystems of the meeting analysis system.

One of the early research studies in observing human activities in an instrumented room is described in [129]. A graphical summary of the human activity is generated. The audio and visual information is used in identifying the current speaker based on a rule based decision fusion. [130] and [131]

TABLE IV
SUMMARY OF STREAM WEIGHT SELECTION SCHEMES IN AUDIO-VISUAL SPEECH RECOGNITION

Stream weight selection criterion	Model	Publication and Year
Empirically determined linear function of the SNR	Spectrum based classification	Yuhas et. al. [71] 1989
Neural network weights trained to minimum error	Time delay neural network (TDNN)	Stork et. al. [91] 1992
Relative entropy - between all audio and video phone state activations	Multi state Time delay neural network (MS-TDNN)	Bregler et. al. [72] 1993
N-best word probability dispersion	Hidden Markov Model (HMM)	Adjoudani and Benoit [73] 1996
No weighting used - performs worse than with weighting	Semi-continuous HMM	Su and Silsbee [92] 1996
Probabilistic gradient descent on classification error	Multistream HMM	Potamianos and Graf [79] 1998
Probabilistic gradient descent to maximize the relative likelihood of the best word	Multistream HMM	Nakamura et. al. [80] 2000
Max. Entropy and Min. classification error	Multistream HMM	Gravier et. al. [81] 2002
Empirically determined for given SNR	Coupled HMM	Nefian et. al. [82] 2002
N-best log likelihood difference and N-best log likelihood dispersion	Multistream HMM	Potamianos et. al. [16] 2003
AdaBoost	Boosted HMM classifier	Yin et. al. [86] 2003
Minimum word error after reducing to two stream	Dynamic Bayesian Networks	Gowdy et. al. [85] 2004
Empirically determined for given SNR	Hybrid SVM-HMM	Gurban and Thiran [87] 2005
Stream entropy	Multistream HMM	Gurban et. al. [42] 2008
Variance of extrinsic information	Multistream HMM	Shivappa et. al. [33] 2008

describe another meeting room analysis system which also fuses audio-visual stream for person identification, in addition to using the audio for automatic transcription and archival purposes. [132] investigates speech, gaze and gesture cues for high level segmentation of a discourse into topical segments based on a psycholinguistic model.

More recent work in [24] models the action of the group of individuals in a meeting instead of individual actions. HMMs are used to statistically model the state of the group using audio and video features and the interactions between individuals are inherently accounted for in the model. Using this formulation, meetings are segmented into five categories: Discussions, Monologues, Note-Taking, Presentations and White-Board presentations. Different fusion schemes were evaluated and the early integration strategy performed the best followed closely by the asynchronous HMM. The feature concatenation scheme could suffer from the curse of dimensionality. Intutively, there is a certain amount of asynchrony between the audio and visual streams in a meeting scene and this hints at the possible inadequacy of using simple HMMs to model the meeting scenes. [133] describes a two layered HMM model to segment the meeting at the individual and group levels respectively. In this case, the asynchronous HMM performs best at the lower level as expected. Dynamic Bayesian networks were explored for suitability in modeling meetings in [134]. An comparison of various modeling techniques is provided in [135].

A number of multimodal meeting rooms equipped with multimodal sensors have been established by various research groups and consortiums. Annotated audio-visual corpora have been collected and standard evaluations have been organized to compare existing frameworks on specific tasks. Table VIII lists the details of a few important meeting corpora. Another recent effort in collecting and organizing multimodal corpora is presented in [136].

Recent evaluations of meeting scene analysis systems include the CLEAR 2006 evaluation [9] and CLEAR 2007 evaluation [10].

V. EXISTING CHALLENGES AND FUTURE DIRECTIONS

Significant progress has been made in the areas of multimodal human activity analysis. However, there exists significant challenges to enable human-like intelligence to intelligent spaces. In this section we discuss two important issues that we consider are central to developing systems capable of human-like intelligence.

TABLE V
SUMMARY OF FUSION STRATEGIES IN AUDIO-VISUAL PERSON LOCALIZATION AND TRACKING

Fusion strategy for audio-visual person localization and tracking	Sensors	Scene complexity	Model	Publication and Year
Proximity based speaker association	1,2	S	Camera epipolar geometry and audio cross-correlation	Pingali et. al. [95] 1999
SNR based weighted average of SPMs	2,3	S	Spatial probability maps	Aarabi [96] 2001
Feature concatenation without weighting	1,2	S	Probabilistic tracking with particle filters	Vermaak et. al. [103] 2001
Proximity based association of audio and visual events	1,4	M	Auditory epipolar geometry and face localization	Nakadai et. al. [97] 2001
Product rule	2,14	M	Probabilistic tracking with particle filters	Zotkin et. al. [104] 2002
Importance sampling and product rule	2,14	M	Probabilistic tracking with particle filters	Gatica-Perez et. al. [105] 2003
Speaker detection using audio	1,3	M	Skin tone based face detection in omni-camera	Kapralos et. al. [98] 2003
Feature concatenation without weighting	1,2	M	Bayesian network	Beal et. al. [66] 2003
Weighted addition of proposal distributions from each sensor	5,2	M	Probabilistic tracking with particle filters	Chen and Rui [64] 2004
Product rule	2,16	M	Probabilistic tracking with particle filters	Checka et. al. [63] 2004
Feature concatenation without weighting	4,12	M	Probabilistic tracking with particle filters	Nickel et. al. [38] 2005
Sequential state update using audio and video	4,16	M	Iterated extended Kalman filter	Gehrig et. al. [26] 2005
Feature concatenation without weighting	2,14	M	Markov Chain Monte Carlo particle filter	Gatica-Perez et. al. [62] 2007
Feature concatenation without weighting	4,14	M	Particle filter	Bernardin et. al. [39] 2007
Finite state machine for appropriate weighting	1,14	M	Particle filter	Bernardin et. al. [39] 2007
Iterative decoding algorithm	2, 8	M	Hidden Markov Model	Shivappa et. al. [99] 2010

A. Hierarchical Fusion Strategies

While little is known on how humans understand and interpret the complex world, the consensus is that an integration of information at different levels of the semantic hierarchy has to come together for this task. Most recently, the researchers in the intelligent systems design have also started exploring hierarchical fusion schemes. In practice, an intelligent space or system equipped with multiple audio-visual sensors, can extract many kinds of audio and video cues such as sound source location, person tracks, speech content, beamformed or enhanced speech, speaker identity etc. Significant benefits of fusion emerge from using the different audio and visual cues together. For example, in [143], the audio localization is used to focus the attention of the video processor which uses skin tone detection and face detection to interact with the user. Such integration is necessary to design and develop context aware systems [144]. In [17], the authors develop a hierarchical fusion framework and explore the relationship between tasks such as person tracking, speech recognition, beamforming, speaker identification, head pose estimation and key word spotting. It is demonstrated that these tasks can be synergetically performed and the whole is greater than the sum of the parts. [145] outlines a meeting analysis task that is based on the probabilistic fusion of audio and visual cues. In [146], the authors propose a hierarchial HMM framework for modeling human activity. More recent hierarchical fusion strategies include [133][147][148][149]. In [149], the authors develop a probabilistic integration framework for fusion of audio visual cues at the track and identity levels. This is an example of fusion at multiple levels of abstraction. Similarly, in [18], the utility of head pose estimation and tracking for speech recognition from distant microphones is explored. In [19], the authors use video localization to enhance the performance of the beamformer for better speech reconstruction from far field microphones. The utility of hierarchical fusion to develop robust human activity analysis algorithms is quite

TABLE VI
SUMMARY OF FUSION STRATEGIES IN AUDIO-VISUAL EMOTIONAL STATE RECOGNITION

Fusion strategy for audio-visual emotion detection tasks	Language	Model(Video)	Model(Audio)	Publication and Year
perceptually dominant modality based weighting	Spanish, Sinhala	—	—	De Silva et. al. [113] 1997
Decision tree-like classifier on audio and video features	Spanish, Sinhala	—	—	Chen et. al. [106] 1998
Dominant mode prevails in case of conflict	Sinhala	Nearest neighbor	HMM	De Silva and Ng [107] 2000
Weighted average of confidence scores	Japanese	Neural Network	HMM	Yoshitomi et. al. [108] 2000
—	Am. English	Sparse network of winnows	Gaussian density	Chen and Huang [109] 2000
Emotion with maximum score among audio and video scores is chosen	Korean	Wavelets and LDA	Wavelets and codebook based multiple band classification	Go et. al. [114] 2003
Feature fusion and decision fusion	Am. English	SVM	SVM	Busso et. al. [110] 2004
Fisher boosting classifier	Am. English	—	—	Zeng et. al. [47] 2005
Weighted average of confidence scores	German and Am. English	SVM and ANN	SVM	Hoch et. al. [111] 2005
Bayesian networks	—	—	—	Sebe et. al. [112] 2006
Multistream Fused HMM	Am. English	HMM	HMM	Zeng et. al. [115] 2008
Decision fusion using Bayesian networks	Am. English	GMM and HMM	HMM	Metallinou et. al. [116] 2010

evident from these existing examples.

In Figure 8, we present a flow diagram of the fusion of multimodal cues explored in [17]. The audio and video signals provide the person location information and this is fused in the audio-visual tracking step to come up with robust estimates of the 3D co-ordinates of the subjects. The tracking information is augmented with the speaker ID when available and this betters the re-identification of the tracks in ambiguous cases. The location and head pose estimates are fused for effective beamforming. The reconstructed clean speech from the beamformer is used by the speaker ID module which identifies the active speaker. The speech recognizer uses both the speaker ID and the reconstructed speech to recognize full speech or spot keywords in the utterance.

Thus, when the various blocks for audio-visual human activity analysis are put together, there is a whole range of fusion possibilities to make the system more robust and effective. As an example, one can consider the fusion of head dynamics, gestures and speech that is explored in [150]. This comprehensive fusion hierarchy, combining audio and visual cues at such varying levels of abstractions to achieve a set of tasks together is an important research direction and there is a need to develop a formal probabilistic framework to address the same. Though the benefits of such hierarchical fusion schemes are quite evident, the choice of cues to use and the fusion framework is very domain specific. There is a great interest in developing adaptive frameworks that can learn and adapt to new scenes as well as sensor configuration. In the next section, we explore this issue in detail.

B. Learning from Multimodal Correspondences

Learning in audio-visual systems can involve various aspects of the framework. Significant among them are

- Automatic calibration for learning new sensor configurations.
- Unsupervised learning for model adaptation in new scenes.
- Model learning - learning the relevant correspondences similar to human learning based on cognitive studies.

C. Automatic calibration

In this survey we have presented several systems that use multiple audio and video sensors. A major issue in deploying these systems is the calibration of the sensors with respect to each other and with respect to the world co-ordinates. Any change in the position or the orientation of the sensors leads to the system having to be re-calibrated. One way to avoid

TABLE VII
SUMMARY OF FUSION STRATEGIES IN AUDIO-VISUAL PERSON IDENTIFICATION

Fusion strategy for audio-visual person identification	Num. Sub-jects	Model(Video)	Model(Audio)	Publication and Year
Empirically determined weighted combination of classifier scores	10	ANN	ANN	Chibelushi et. al. [123] 1993
Weighted average of scores	89	Nearest neighbor	Vector quantization	Brunelli and Falavigna[124] 1995
Weighted average of scores	37	HMM	HMM	Jourlin et. al. [125] 1997
Weighted average of scores based on classification error	37	GMM	GMM	Wark et. al. [126] 1999
SVM	295	Robust correlation	Harmonic sphericity based similarity score	Ben-Yacoub et. al. [127] 1999
Adaptive weighted combination of scores based on dispersion confidence measure	37	GMM	GMM	Wark and Sridharan [120]2001
Adaptive cascade with the ordering based on reliability of classifier	50	Eigen faces	HMM	Erzin et. al. [128] 2005
Reliability -weighted summation of scores	50	GMM	GMM	Erzin et. al. [34] 2006

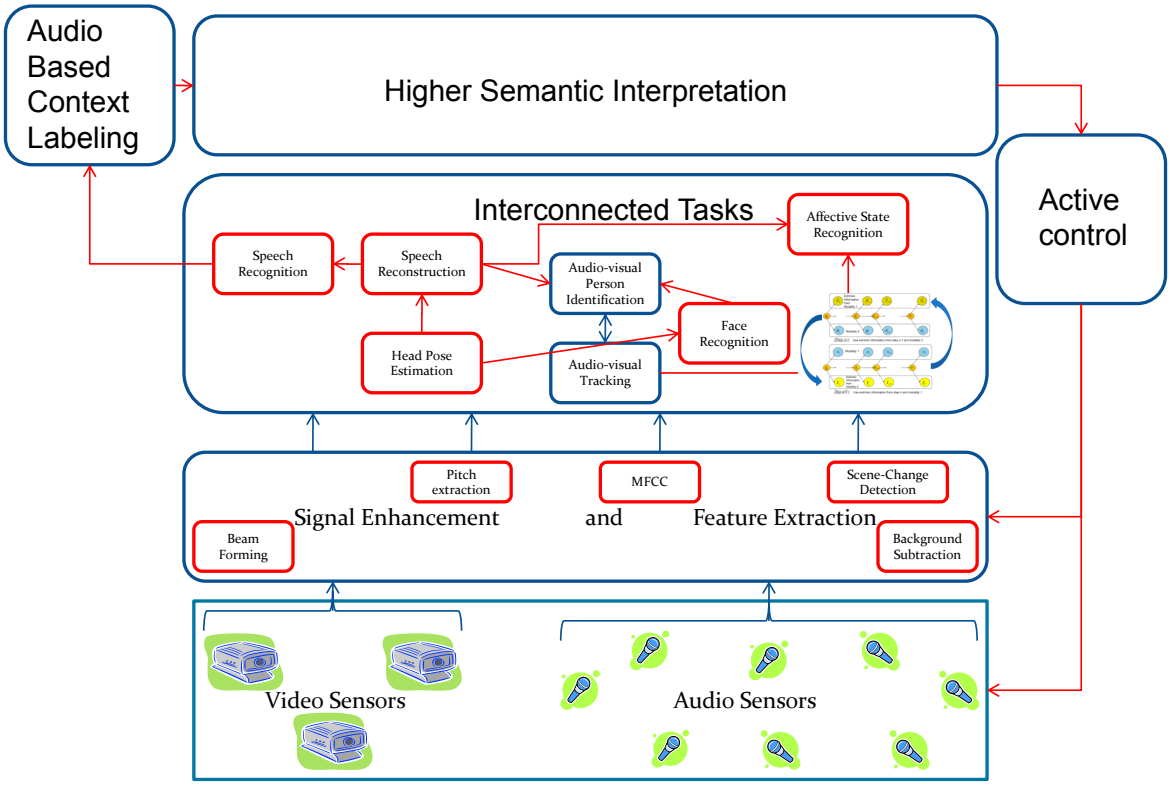


Fig. 8. The overall organization of a general human activity analysis in audio-visual spaces [17].

this issue is to develop fusion algorithms that do away with the accurate calibration and can learn the correspondences between sensors while being operational. [99] and [100] are examples of such algorithms. Audio-visual systems also allow for evolving new ways to calibrate sensors as well. Also popular in literature are systems which have fixed geometry and hence can be pre-calibrated as in [25]. Mounting the sensors on the users is also explored in certain applications such as gaming and military applications. In any case, the sensor calibration issue needs to be resolved to facilitate the

TABLE VIII

STANDARD AUDIO-VISUAL MEETING SCENE CORPORA AND THEIR SENSORY/SCENE/PARTICIPANT INFORMATION.

ISL [137] : The Interactive System Labs of CMU, Pittsburgh has collected a database consisting of more than 100 diverse meetings, combined total of 103 hours (4.3 days). Each meeting lasted an average of 60 minutes. The meetings have an average of 6.4 participants. The meetings have been collected since 1999. A meeting in the database is a minimum of three individuals speaking to one another. The results are presented in a maximum of eight mono audio files in WAV format, so-called speaker and recording protocol files containing information about the participants, equipment, environment and scenario, three video tapes, one transcription file of the entire meeting, so-called marker file containing begin and end time stamps for conversation contributions, and a list of the meetings vocabulary. The meeting scenarios include ProjectWork Planning, Military Block Parties, Games, Chatting, and Topic Discussion.

ICSI [138] : International Computer Science Institute, Berkeley, California has collected a 75-meeting corpus with audio and transcripts of natural meetings recorded simultaneously with head-worn and tabletop microphones. The corpus contains 75 meetings of 4 main types and 53 unique speakers. The data totals to over 70 meeting-hours and up to 16 channels for each meeting. The ICSI effort is predominantly an audio scene analysis and meeting transcription effort.

NIST [139] : NIST has constructed a Meeting Data Collection Laboratory (MDCL) to collect corpora to support meeting domain research, development and evaluation. The NIST Smart Data Flow architecture, developed by the NIST Smart Spaces Laboratory, streams and captures all of the sensor data from 200 mics and 5 video cameras on 9 separate data collection systems in a proprietary time-indexed SMD format. This architecture also ensures that all data streams are synchronized (via the Network Time Protocol and NIST atomic clock signal) to within a few milliseconds. The NIST Meeting Room Pilot Corpus consists of 19 meetings/15 hours recorded between 2001 and 2003. In total, the multi-sensor data comes to 266 hours of audio and 77 hours of video.

CHIL [140] : The CHIL (Computers in the Human Interaction Loop) consortium is an European reserach effort with the participation of 15 partner sites from nine countries under the joint coordination of the Fraunhofer - IITB and the Interactive Systems Labs (ISL) of the University of Karlsruhe, Germany. Five smart rooms have been set up as part of the CHIL project, and have been utilized in the data collection efforts. Two types of interaction scenarios constitute the focus of the CHIL corpus: lectures and meetings. The CHIL corpus is accompanied by rich manual annotations of both its audio and visual modalities. In particular, it contains a detailed multi-channel verbatim orthographic transcription of the audio modality that includes speaker turns and identities, acoustic condition information, and name entities for part of the corpus. Furthermore, video labels provide multi-person head location in the 3D space, as well as information about the 2D face bounding box and facial feature locations visible in all camera views. In addition, head-pose information is provided for part of the corpus. Each smart room contains a minimum of 88 microphones that capture both close-talking and far-field acoustic data. There exists at least one 64-channel linear microphone array, namely the Mark III array developed by NIST. The video data is captured by five fixed cameras. Four of them are mounted close to the corners of the room, by the ceiling, with significantly overlapping and wide-angle fields-of-view.

use of audio-visual systems in practice.

D. Unsupervised learning

Significant progress has been made in designing systems that fuse audio and visual information to achieve better accuracy and robustness to background noise. A very commonly seen paradigm in fusion schemes is the appropriate weighting of input streams according to their reliability. The assessment of the quality of individual streams is a challenging task in itself and needs further research. Common measures like SNR are useful but there is a necessity for other measures to quantify the reliability of extracted cues from audio and video streams. As an example, consider a speaker whose lips are sometimes partially or fully occluded from the camera due to his changing orientation on an audio-visual speech recognition system. How to enable the system to weight the audio and visual cues appropriately in this case?

The question can be reposed as, how to build a system that can adapt to changing situations? This leads us to the bigger problem of learning. Most of the systems described in this survey learn model parameters in a supervised fashion. This requires a lot of annotated training data and also places the restriction that the conditions during the deployment of the system cannot be significantly different from the training conditions. This highlights the utility of semi-supervised and unsupervised methods for learning parameters. Semi-supervised learning allows one to learn model parameters from a small amount of annotated training data and large amounts of non-annotated training data. Unsupervised learning has also been used in ambient intelligence systems to classify previously unseen activities [58]. Existing research has addressed the problem of learning at several levels. Multimodality has an advantage in unsupervised learning through the presence of cross-modal correspondences.

TABLE IX
STANDARD AUDIO-VISUAL MEETING SCENE CORPORA AND THEIR SENSORY/SCENE/PARTICIPANT INFORMATION. (CONTD.)

VACE [141] : Under this research effort, Air Force Institute of Technology (AFIT) modified a lecture room to collect multimodal, time-synchronized audio, video, and motion data. In the middle of the room, up to 8 participants can sit around a rectangular conference table. 10 camcorders and 9 Vicon MCam2 near-IR cameras, driven by the Vicon V8i Data Station record the video data. For audio, the participants wear Countryman ISOMAX Earset wireless microphones to record their individual sound tracks. Table-mounted wired microphones are used to record the audio of all participants (two to six XLR-3M connector microphones configured for the number of participants and scenario, including two cardioid Shure MX412 D/C microphones and several types of low-profile boundary microphones (two hemispherical polar pattern Crown PZM-6D, one omni-directional Audio Technica AT841a, and one four-channel cardioid Audio Technica AT854R). For the VACE meeting corpus, each participant is recorded with a stereo calibrated camera pair. The Vicon system is used to obtain more accurate tracking results to inform subsequent coding efforts, while also providing ground truth for video-tracking algorithms.

AMI & AMIDA [142] : The AMI and AMIDA projects are EU projects concerned with the recognition and interpretation of multiparty meetings.Three standardized meeting rooms were constructed at IDIAP, TNO and University of Edinburgh. Each room consisted of at least 6 cameras and 12 microphones. The different recording streams are synchronized to a common timeline. The corpus consists of 100 hour annotated corpus of meetings, with speech annotations aligned to the word level. Also, manual annotations of the behavior of the meeting participants are provided at various levels namely dialogue acts, topic segmentation, extractive and abstractive summaries, named entitie, gaze direction etc.

TABLE X
SUMMARY OF HIERARCHICAL FUSION STRATEGIES IN AUDIO-VISUAL HUMAN ACTIVITY ANALYSIS

Audio-visual tasks involved	Publication and Year
Human activity recognition	Oliver et. al. [146] 2002
Group and individual activity recognition	Zhang et. al. [133] 2004
Speech Reconstruction - Person Tracking, Beamforming, Speech Recognition	Maganti et. al. [19] 2007
Assitive meeting - Person tracking, Hand tracking, Speaker Orientation and Head pose	Dai and Xu [148] 2008
Identity tracking - Person Tracking, Face recognition, Speaker ID	Bernardin et. al. [149] 2008
Scene Understanding - Person Tracking, Head pose, Beamforming, Speaker ID and Keyword spotting	Shivappa et. al. [17] 2009

E. Model learning

Recent work in cognitive sciences has led to the design of systems that can learn primitive correspondences across different modalities in a manner similar to the learning experiences of a human child. More specifically, systems that ground language in perceptual cues have been proposed. From the previous sections we can conclude that there are a very large number of strategies for fusion of information in human activity analysis systems. Humans are extremely competent at such tasks and seem to employ an near-optimal fusion strategy for each situation. However, most approaches to automatically recognize multimodal actions are based on having a annotated training set[5]. To quote the authors,

> ... However, no matter based on feature or semantic fusion, most systems do not have learning ability in the sense that developers need to encode knowledge into some symbolic representations or probabilistic models during the training phase. Once the systems are trained, they are not able to automatically gain additional knowledge even though they are situated in physical environments and can obtain multisensory information. ...

Yu and Ballard[151] present a unified framework to learn perceptually grounded meanings of spoken words without transcriptions. This is the first step towards building a system that can learn for perceptual cues without the necessity to encode the knowledge in some symbolic representation. The perceptually grounded words provide the symbolic representation[5]. The opposite process where visually-guided attention helps in understanding a complex auditory scenes has also been studied in literature [152].

Modeling schemes influence the fusion strategy used and the modeling schemes are themselves are heavily task oriented, as seen in the preference of the speech recognition community in using HMMs and the tracking community in using particle filters. An intelligent system will have to simultaneously perform these tasks in order to perform tasks like a human. A framework to fuse the different systems would have to be developed. Learning such a framework by starting with a certain amount of pre-programmed intelligence, but streamlining the models and adding extra functionalities both by supervised learning and observing multimodal data for cross-modal correspondences in a manner similar to the development of human cognition is a challenge towards which

the research community is making advances towards.

VI. Concluding Remarks

We have presented an organized review of the various audio-visual fusion schemes. We have presented a systematic organization of the fusion schemes used in audio-visual human activity analysis. We have compared the relative advantages and performances of the various fusion schemes in a task specific manner. We have also discussed in detail the important future steps for developing an audio-visual system capable of human-like behavior by reviewing established and upcoming work in hierarchical fusion strategies and crossmodal learning techniques. We wish to provide the reader a comprehensive view of the challenges and solutions for designing fusion schemes for observing humans with audio and video sensors. To this end, we hope to have filled in the need for a survey dedicated entirely to audio-visual information fusion strategies.

Acknowledgment

We would like to thank our main sponsors, CALIT2 at UC San Diego, NSF's RESCUE project and the UC Discovery program. We also like to acknowledge the assistance provided by our colleagues for reviewing the manuscript and providing valuable suggestions to improve the presentation of the paper. We sincerely thank the reviewers for their valuable advise which has helped us enhance the content as well as the presentation of the paper.

References

[1] A. M. Turing, "Computing machinery and intelligence," *Mind*, 1950.

[2] T. Chen, "Audiovisual speech processing," *IEEE Signal Processing Magazine*, Jan 2001.

[3] D. Roy, "Grounding words in perception and action: computational insights," *Trends in Cognitive Sciences*, Aug. 2005.

[4] G. O. Dek, M. S. Bartlett, and T. Jebara, "New trends in cognitive science: Integrative approaches to learning and development," *Neurocomputing*, Jan. 2007.

[5] C. Yu and D. H. Ballard, "A multimodal learning interface for grounding spoken language in sensory perceptions," *ACM Trans. Appl. Percept.*, vol. 1, no. 1, 2004.

[6] I. Bloch, "Information combination operators for data fusion: a comparative review with classification," *IEEE Transactions on Systems, Man and Cybernetics, Part A*, Jan 1996.

[7] A. Jaimes and N. Sebe, "Multimodal human-computer interaction: A survey," *Comput. Vis. Image Underst.*, 2007.

[8] J. Bijhold, A. Ruifrok, M. Jessen, Z. Geradts, and S. Ehrhardt, "Forensic audio and visual evidence 2004–2007: A review," in *15th INTERPOL Forensic Science Symposium*, Lyon, France, October, 2007.

[9] R. Stiefelhagen, K. Bernardin, R. Bowers, J. S. Garofolo, D. Mostefa, and P. Soundararajan, "The clear 2006 evaluation," in *CLEAR*.

[10] R. Stiefelhagen, R. Bowers, and J. Fiscus, *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007 (Lecture Notes in Computer Science)*.

[11] B. Kane, S. Luz, and J. Su, "Capturing multimodal interaction at medical meetings in a hospital setting: Opportunities and challenges," in *LREC 2010 workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*.

[12] A. Fleury, M. Vacher, F. Portet, P. Chahuara, and N. Noury, "A multimodal corpus recorded in a health smart home," in *LREC 2010 workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*.

[13] C. Tran and M. M. Trivedi, "Towards a vision-based system exploring 3d driver posture dynamics for driver assistance: Issues and possibilities," in *IEEE Intelligent Vehicles Symposium*, 2010.

[14] A. Tawari and M. M. Trivedi, "Contextual framework for speech based emotion recognition in driver assistance system," in *IEEE Intelligent Vehicles Symposium*, 2010.

[15] K. Takeda, H. Erdogan, J. H. L. Hansen, and H. Abut, *In-Vehicle Corpus and Signal Processing for Driver Behavior (Springer USA 2009)*.

[16] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, Sept. 2003.

[17] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Hierarchical audio-visual cue integration framework for activity analysis in intelligent meeting rooms," in *IEEE CVPR Workshop: ViSU'09*, 2009.

[18] S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "Role of head pose estimation in speech acquisition from distant microphones," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

[19] H. K. Maganti, D. Gatica-Perez, and I. McCowan, "Speech enhancement and recognition in meetings with an audio-visual sensor array," *IEEE Trans. on Audio, Speech, and Language Processing*, Nov 2007.

[20] P. Teissier, J. Robert-Ribes, J. L. Schwartz, and A. Guerin-Dugue, "Comparing models for audiovisual fusion in a noisy-vowel recognition task," *IEEE Transactions on Speech and Audio Processing*, Nov 1999.

[21] J. L. Schwartz, J. Robert-Ribes, and P. Escudier, "Ten years after summerfield: A taxonomy of models for audio-visual fusion in speech perception," *Hearing by Eye II*, 1998.

[22] A. M. Martinez and A. C. Kak, "Pca versus lda," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Feb 2001.

[23] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri, "Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins summer 2000 workshop," in *Proceedings of IEEE Workshop Multimedia Signal Processing*, 2001.

[24] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Mar. 2005.

[25] A. ODonovan and R. Duraiswami, "Microphone arrays as generalized cameras for integrated audio visual processing," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2007.

[26] T. Gehrig, K. Nickel, H. K. Ekenel, U. Klee, and J. McDonough, "Kalman filters for audio-video source localization," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*, Oct. 2005.

[27] A. Garg, G. Potamianos, C. Neti, and T. S. Huang, "Frame-dependent multi-stream reliability indicators for audio-visual speech recognition," in *Proceedings of the International Conference on Multimedia and Expo*, 2003.

[28] K. W. Grant and S. Greenberg, "Speech intelligibility derived from asynchronous processing of auditory-visual information," 2001.

[29] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, Sept. 2000.

[30] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust asr," *Speech Communication*, 2001.

[31] A. P. Varga and R. K. Moore, "Hidden markov model decomposition of speech and noise," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Apr 1990.

[32] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[33] S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.

[34] E. Erzin, Y. Yemez, A. M. Tekalp, A. Ercil, H. Erdogan, and H. Abut, "Multimodal person recognition for human-vehicle interaction," *IEEE Multimedia Magazine*, 2006.

[35] C. Canton-Ferrer, C. Segura, J. R. Casas, M. Pardàs, and J. Hernando, "Audiovisual head orientation estimation with particle filtering in multisensor scenarios," *EURASIP J. Adv. Signal Process*, vol. 2008.

[36] R. Stiefelhagen, H. K. Ekenel, C. Fugen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, and A. Waibel, "Enabling multimodal humanrobot interaction for the karlsruhe humanoid robot," *IEEE Transactions on Robotics*, Oct. 2007.

[37] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Person tracking with audio-visual cues using the iterative decoding framework," in *5th*

IEEE International Conference On Advanced Video and Signal Based Surveillance, 2008 [Best Paper Award].

[38] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, "A joint particle filter for audio-visual speaker tracking," in *Proceedings of the 7th international conference on multimodal interfaces*, 2005.

[39] K. Bernardin, T. Gehrig, and R. Stiefelhagen, "Multi-Level Particle Filter Fusion of Features and Cues for Audio-Visual Person Tracking," in *CLEAR Evaluation Workshop*, 2007.

[40] W. Zajdel, J. Krijnders, T. Andringa, and D. Gavrila, "Cassandra: audio-video sensor fusion for aggression detection," in *4th IEEE International Conference On Advanced Video and Signal Based Surveillance*, 2007.

[41] S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "An iterative decoding algorithm for fusion of multi-modal information," EURASIP *Journal on Advances in Signal Processing - Special Issue on Human-Activity Analysis in Multimedia Data*, 2008.

[42] M. Gurban, J.-P. Thiran, T. Drugman, and T. Dutoit, "Dynamic modality weighting for multi-stream HMMs in Audio-Visual Speech Recognition," in *10th International Conference on Multimodal Interfaces*, 2008.

[43] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding:turbo-codes," in *Proceedings of the IEEE International Conference on Communications*, May 1993.

[44] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, Nov 2006.

[45] M. Liu, H. Tang, H. Ning, and T. S. Huang, "Person identification based on multichannel and multimodality fusion," in *CLEAR*, 2006.

[46] A. Stergiou, A. Pnevmatikakis, and L. Polymenakos, "A decision fusion system across time and classifiers for audio-visual person identification," in *CLEAR*, 2006.

[47] Z. Zeng, Z. Zhang, B. Pianfetti, J. Tu, and T. S. Huang, "Audio-visual affect recognition in activation-evaluation space," *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, July 2005.

[48] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audio-visual synchronization and fusion using canonical correlation analysis," *IEEE Transactions on Signal Processing*, 2007.

[49] Y. Keller, S. Lafon, R. Coifman, and S. Zucker, "Audio-visual group recognition using diffusion maps," *IEEE Transactions on Signal Processing*, 2009.

[50] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, 2002.

[51] A. Nijholt and G. Hondorp, "Towards communicating agents and avatars in virtual worlds," in *Proceedings Eurographics*, 2000.

[52] A. Nijholt, "Towards multi-modal interactions in virtual environments: A case study," in *Actas-I, VI Simposio Internacional de Communicacion Social*, 1999.

[53] T. S. Huang, M. A. Hasegawa-Johnson, S. M. Chu, and Z. Zeng, "Sensitive talking heads," *Image Vision Computing*, 2009.

[54] C. Breazeal, "Designing sociable robots," *Robotics and Autonomous Systems*, 2002.

[55] C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati, "Active vision for sociable robots," *IEEE Transactions on Systems, Man and Cybernetics, Part A*, Sep 2001.

[56] F. R. M. Baldauf, S. Dustdar, "A survey on context-aware systems," *International Journal of Ad Hoc and Ubiquitous Computing*, 2007.

[57] M. M. Trivedi, K. S. Huang, and I. Mikic, "Dynamic context capture and distributed video arrays for intelligent spaces," *IEEE Transactions on Systems, Man and Cybernetics*, Jan. 2005.

[58] O. Brdiczka, J. Maisonnasse, P. Reignier, and J. Crowley, "Detecting small group activities from multimodal observations," *Applied Intelligence*, 2007.

[59] O. Brdiczka, J. Maisonnasse, and P. Reignier, "Automatic detection of interaction groups," in *Proceedings of the 7th international conference on Multimodal interfaces*, 2005.

[60] D. Perzanowski, A. C. Schultz, and W. Adams, "Integrating natural language and gesture in a robotics domain," *Proceedings of the IEEE International Symposium on Computational Intelligence in Robotics and Automation*, Sep 1998.

[61] S. S. Ghidary, Y. Nakata, H. Saito, M. Hattori, and T. Takamori, "Multi-modal human robot interaction for map generation," *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2001.

[62] D. Gatica-Perez, G. Lathoud, J. Odobez, and I. McCowan, "Audio-visual probabilistic tracking of multiple speakers in meetings," *IEEE Transactions on Audio, Speech and Language Processing*, 2007.

[63] N. Checka, K. W. Wilson, M. R. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.

[64] Y. Chen and Y. Rui, "Real-time speaker tracking using particle filter sensor fusion," *Proceedings of the IEEE*, 2004.

[65] R. Cutler and L. S. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *IEEE International Conference on Multimedia and Expo (III)*, 2000.

[66] M. Beal, N. Jojic, and H. Attias, "A graphical model for audiovisual object tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2003.

[67] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Letters to Nature*, December 1976.

[68] D. G. Stork and M. E. H. eds., *Speechreading by machines and humans*. Berlin, Germany: Springer, 1996.

[69] E. D. Petajan, "Automatic lipreading to enhance speech recognition (speech reading)," Ph.D. dissertation, Champaign, IL, USA, 1984.

[70] E. Petajan, B. Bischoff, D. Bodoff, and N. M. Brooke, "An improved automatic lipreading system to enhance speech recognition," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1988.

[71] B. Yuhas, M. Goldstein, and T. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Communications Magazine*, November 1989.

[72] C. Bregler, S. Manke, and A. Waibel, "Bimodal sensor integration on the example of speech–reading," in *Proc. of IEEE Int. Conf. on Neural Networks*, 1993.

[73] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in an hmm-based asr," *Speechreading by Humans and Machines.*, 1996.

[74] J. Luettin, N. A. Thacker, and S. W. Beet, "Speechreading using shape and intensity information," *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, Oct 1996.

[75] G. Potamianos, E. Cosatto, H. Graf, and D. Roe, "Speaker independent audio-visual database for bimodal asr," in *Proc. Europ. Tut. Work. Audio-Visual Speech Proc.*, 1997.

[76] L. Rabiner and B. Juang, *Fundamentals of speech recognition*. NJ, USA: Prentice-Hall, Inc., 1993.

[77] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. ICSLP*, 1996.

[78] P. Beyerlein, "Discriminative model combination," *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 1998.

[79] G. Potamianos and H. Graf, "Discriminative training of hmm stream exponents for audio-visual speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1998.

[80] S. Nakamura, H. Ito, and K. Shikano, "Stream weight optimization of speech and lip image sequence for audio-visual speech recognition," in *Proc. ICSLP*, 2000.

[81] G. Gravier, S. Axelrod, G. Potamianos, and C. Neti, "Maximum entropy and mce based hmm stream weight estimation for audio-visual asr," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[82] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, and K. Murphy, "Dynamic bayesian networks for audio-visual speech recognition," *EURASIP J. Appl. Signal Process.*, vol. 2002.

[83] G. G. Zweig, "Speech recognition with dynamic bayesian networks," Ph.D. dissertation, Berkeley, CA, USA, 1998.

[84] T. Stephenson, H. Bourlard, S. Bengio, and A. Morris, "Automatic speech recognition using dynamic bayesian networks with both acoustic and articulatory variables," in *Proc. ICSLP vol. II*, 2000.

[85] J. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes, "Dbn based multi-stream models for audio-visual speech recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2004.

[86] P. Yin, I. Essa, and J. M. Rehg, "Boosted audio-visual hmm for speech reading," *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, Oct. 2003.

[87] M. Gurban and J. Thiran, "Audio-Visual Speech Recognition with a Hybrid SVM-HMM System," in *13th European Signal Processing Conference (EUSIPCO)*, 2005.

[88] K. Wilson, V. Rangarajan, N. Checka, and T. Darrell, "Audiovisual arrays for untethered spoken interfaces," *Fourth IEEE International Conference on Multimodal Interfaces*, 2002.

[89] F. Asano, K. Yamamoto, I. Hara, J. Ogata, T. Yoshimura, Y. Motomura, N. Ichimura, and H. Asoh, "Detection and separation of speech event

using audio and video information fusion and its application to robust speech interface," *EURASIP J. Appl. Signal Process.*, 2004.

[90] T. Gustafsson, B. D. Rao, and M. M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Transactions on Speech and Audio Processing*, Nov. 2003.

[91] D. G. Stork, G. Wolff, and E. Levine, "Neural network lipreading system for improved speech recognition," *International Joint Conference on Neural Networks*, Jun 1992.

[92] Q. Su and P. L. Silsbee, "Robust audiovisual integration using semicontinuous hidden markov models," in *Proc. ICSLP*, 1996.

[93] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, 2006.

[94] J. H. DiBiase, H. F. Silverman, and M. S. Branstein, "Robust localization in reverberant rooms," *Microphone Arrays: Signal Processing Techniques and Applications*, 2001.

[95] G. Pingali, G. Tunali, and I. Carlbom, "Audio-visual tracking for natural interactivity," in *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, 1999.

[96] S. G. Z. P. Aarabi, "Robust sound localization using multi-source audiovisual information fusion," *Information Fusion*, 2001.

[97] K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano, "Real-time auditory and visual multiple-object tracking for humanoids," in *IJCAI*, 2001.

[98] B. Kapralos, M. R. M. Jenkin, and E. Milios, "Audiovisual localization of multiple speakers in a video teleconferencing setting," 2003.

[99] S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "Audio visual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation," *IEEE Journal of Selected Topics in Signal Processing (to appear)*, 2010.

[100] E. Ettinger and Y. Freund, "Coordinate-free calibration of an acoustically driven camera pointing system," in *International Conference on Distributed Smart Cameras (ICDSC)*, 2008.

[101] J. W. Fisher, T. Darrell, W. T. Freeman, and P. A. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *NIPS*, 2000.

[102] J. Hershey and J. Movellan, "Audio vision: Using audiovisual synchrony to locate sounds," in *NIPS*, 2000.

[103] J. Vermaak, M. Gangnet, A. Blake, and P. Perez, "Sequential monte carlo fusion of sound and vision for speaker tracking," *Eighth IEEE International Conference on Computer Vision*.

[104] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP J. Appl. Signal Process.*, 2002.

[105] D. G. Perez, G. Lathoud, I. McCowan, J. M. Odobez, and D. Moore, "Audio-visual speaker tracking with importance particle filters," 2003.

[106] L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu, "Multimodal human emotion/expression recognition," *Third IEEE International Conference on Automatic Face and Gesture Recognition,*, Apr 1998.

[107] L. C. D. Silva and P. C. Ng, "Bimodal emotion recognition," *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[108] Y. Yoshitomi, K. Sung-Ill, T. Kawano, and T. Kilazoe, "Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face," *9th IEEE International Workshop on Robot and Human Interactive Communication*, 2000.

[109] L. S. Chen and T. S. Huang, "Emotional expressions in audiovisual human computer interaction," *IEEE International Conference on Multimedia and Expo*, 2000.

[110] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*, 2004.

[111] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll, "Bimodal fusion of emotional data in an automotive environment," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2005.

[112] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Emotion recognition based on joint visual and audio cues," in *Proceedings of the 18th International Conference on Pattern Recognition*, 2006.

[113] L. C. D. Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," *Proceedings of the International Conference on Information, Communications and Signal Processing*, Sep 1997.

[114] H. Go, K. Kwak, D. Lee, and M. Chun, "Emotion recognition from the facial image and speech signal," *SICE Annual Conference*, Aug. 2003.

[115] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang, "Audiovisual affective expression recognition through multistream fused hmm," *IEEE Transactions on Multimedia*, June 2008.

[116] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010.

[117] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, Jan. 2004.

[118] P. S. Aleksic and A. K. Katsaggelos, "Audio-visual biometrics," *Proceedings of the IEEE*, Nov. 2006.

[119] G. Chetty and M. Wagner, "Biometric person authentication with liveness detection based on audio visual fusion," *Int. Journal of Biometrics*, 2009.

[120] T. Wark and S. Sridharan, "Adaptive fusion of speech and lip information for robust speaker identification," *Digital Signal Processing*, 2001.

[121] C. Sanderson and K. Paliwal, "Identity verification using speech and face information," *Digital Signal Processing*, 2004.

[122] 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 Transactions on Pattern Analysis and Machine Intelligence*, March 2005.

[123] C. Chibelushi, F. Deravi, and J. Mason, "Voice and facial image integration for person recognition," in *Proceedings of the IEEE International Symposium on Multimedia Technologies and Future Applications*, 1993.

[124] R. Brunelli and D. Falavigna, "Person identification using multiple cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Oct 1995.

[125] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Integrating Acoustic and Labial Information for Speaker Identification and Verification," in *Proceedings of the European Conference on Speech Communication and Technology*, 1997.

[126] T. Wark, S. Sridharan, and V. Chandran, "Robust speaker verification via fusion of speech and lip modalities," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Mar 1999.

[127] S. Ben-Yacoub, J. Luettin, K. Jonsson, J. Matas, and J. Kittler, "Audio-visual person verification," *cvpr*, 1999.

[128] E. Erzin, Y. Yemez, and A. M. Tekalp, "Multimodal speaker identification using an adaptive classifier cascade based on modality reliability," *IEEE Transactions on Multimedia*, Oct. 2005.

[129] M. M. Trivedi, K. S. Huang, and I. Mikic, "Activity monitoring and summarization for an intelligent meeting room," in *Proceedings of the IEEE International Workshop on Human Motion*, 2000.

[130] M. Bett, R. Gross, H. Yu, X. Zhu, Y. Pan, J. Yang, and A. Waibel, "Multimodal meeting tracker," in *in Proceedings of RIAO2000*, 2000.

[131] R. Gross, M. Bett, H. Yue, X. J. Zhu, Y. Pan, J. Yang, and A. Waibel, "Towards a multimodal meeting record," 2000.

[132] F. Quek, D. McNeill, R. Bryll, C. Kirbas, H. Arslan, K. E. McCullough, N. Furuyama, and R. Ansari, "Gesture, speech, and gaze cues for discourse segmentation," *IEEE conference on Computer Vision and Pattern Recognition*, 2000.

[133] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud, "Modeling individual and group actions in meetings: A two-layer hmm framework," *IEEE International Conference on Computer Vision and Pattern Recognition*, June 2004.

[134] A. Dielmann and S. Renals, "Automatic meeting segmentation using dynamic bayesian networks," *IEEE Transactions on Multimedia*, Jan. 2007.

[135] M. Al-Hames, C. Lenz, S. Reiter, J. Schenk, F. Wallhoff, and G. Rigoll, "Robust multi-modal group action recognition in meetings from disturbed videos with the asynchronous hidden markov model," *IEEE International Conference on Image Processing*, 2007.

[136] M. Kipp, J. C. Martin, P. Paggio, and D. Heylen, *Multimodal Corpora: From Models of Natural Interaction to Systems and Applications*. Lecture Notes on Aritificial Intelligence, Springer, 2009.

[137] S. Burger, V. MacLaren, and H. Yu, "The isl meeting corpus: the impact of meeting type on speech style," in *ICSLP*, 2002.

[138] N. Morgan, D. Baron, S. Bhagat, H. Carvey, R. Dhillon, J. Edwards, D. Gelbart, A. Janin, A. Krupski, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "Meetings about meetings: research at icsi on speech in multiparty conversations," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, April 2003.

[139] J. Garofolo, C. Laprum, M. Michel, V. Stanford, and E. Tabassi, "The NIST meeting room pilot corpus," in *in Proc. of Language Resource and Evaluation Conference.*, 2004.

[140] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, J. R. Casas, J. Turmo, L. Cristoferetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet, "The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms," *Journal on Language Resources and Evaluation*, Dec 2007.

[141] L. Chen, R. T. Rose, Y. Qiao, I. Kimbara, F. Parrill, T. X. Han, J. Tu, Z. Huang, M. Harper, Y. Xiong, D. Mcneill, R. Tuttle, and T. Huang, "Vace multimodal meeting corpus," in *in Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, 2005.

[142] S. Renals, T. Hain, and H. Bourlard, "Interpretation of multiparty meetings: The ami and amida projects," in *HSCMA*, April 2008.

[143] S. Cheamanunkul, E. Ettinger, M. Jacobsen, P. Lai, and Y. Freund, "Detecting, tracking and interacting with people in a public space," in *Proceedings of the 2009 international conference on Multimodal interfaces, ICMI-MLMI*, 2009.

[144] A. Pnevmatikakis, J. Soldatos, F. Talantzis, and L. Polymenakos, "Robust multimodal audio-visual processing for advanced context awareness in smart spaces," *Personal Ubiquitous Computing*, 2009.

[145] K. Ishizuka, S. Araki, K. Otsuka, T. Nakatani, and M. Fujimoto, "A speaker diarization method based on the probabilistic fusion of audio-visual location information," in *Proceedings of the 2009 international conference on Multimodal interfaces, ICMI-MLMI*, 2009.

[146] N. Oliver, E. Horvitz, and A. Garg, "Layered representations for human activity recognition," in *Proceedings of International Conference on Multimodal Interfaces*, Oct. 2002.

[147] N. M. Oliver, B. Rosario, and A. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[148] P. Dai and G. Xu, "Context-aware computing for assistive meeting system," in *Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments*, 2008.

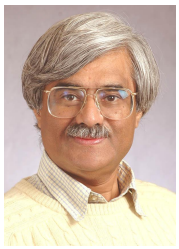
[149] K. Bernardin, R. Stiefelhagen, and A. Waibel, "Probabilistic integration of sparse audio-visual cues for identity tracking," in *Proceeding of the 16th ACM international conference on Multimedia*, 2008.

[150] P. Paggio and C. Navarretta, "Feedback in head gestures and speech," in *LREC 2010 workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*.

[151] C. Yu and D. H. Ballard, "A unified model of early word learning: Integrating statistical and social cues," *Neurocomput.*, 2007.

[152] V. Best, E. J. Ozmeral, and B. G. Shinn-Cunningham, "Visually-guided attention enhances target identification in a complex auditory scene," *Journal of the Association for Research in Otolaryngology*, 2007.

Shankar T. Shivappa Shankar T. Shivappa received his B.Tech. and M.Tech degrees in Electrical Engineering from the Indian Institute of Technology, Madras, India, in 2004. He is currently a Ph.D. candidate in the Department of Electrical and Computer ngineering at UCSD. His research interests lie in the areas of multimodal signal processing, machine learning, Speech and audio processing and computer vision. He has interned at AT&T labs during the summer of 2005 and at Microsoft Research during the summer of 2009. His paper, co-authored with Mohan Trivedi and Bhaskar Rao, received the best paper award at AVSS 2008. He is currently actively involved in the running of the Smart space laboratory at the California Institute for Telecommunication and Information Technologies [Cal-IT2], UCSD.



Mohan M. Trivedi Mohan Manubhai Trivedi(IEEE Fellow) received B.E. (with honors) degree from the Birla Institute of Technology and Science, Pilani, India, and the Ph.D. degree from Utah State University, Logan. He is Professor of electrical and computer engineering and the Founding Director of the Computer Vision and Robotics Research Laboratory, University of California, San Diego (UCSD), La Jolla. He has established the Laboratory for Intelligent and Safe Automobiles (LISA), UCSD, to pursue a multidisciplinary research agenda. He and his team are currently pursuing research in machine and human perception, active machine learning, distributed video systems, multimodal affect and gesture analysis, human- centered interfaces, intelligent driver assistance systems. He regularly serves as a consultant to industry and government agencies in the U.S. and abroad. He has given over 55 keynote/plenary talks. Prof. Trivedi served as the Editor-in-Chief of the Machine Vision and Applications journal. He is currently an editor for the IEEE Transactions on Intelligent Transportation Systems and Image and Vision Computing. He served as the General Chair for IEEE Intelligent Vehicles Symposium IV 2010. He has received the Distinguished Alumnus Award from Utah State University, Pioneer Award and Meritorious Service Award from the IEEE Computer Society, and several Best Paper Awards. Trivedi is a Fellow of the IEEE, IAPR, and the SPIE.

Bhaskar D. Rao Bhaskar D. Rao (IEEE Fellow) received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, India, in 1979 and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1981 and 1983, respectively. Since 1983, he has been with the University of California at San Diego, La Jolla, where he is currently a Professor with the Electrical and Computer Engineering Department and holder of the Ericsson endowed chair in wireless access networks. His interests are in the areas of digital signal processing, estimation theory, and optimization theory, with applications to digital communications, speech signal processing, and human-computer interactions. He is the holder of the Ericsson endowed chair in Wireless Access Networks and is the Director of the Center for Wireless Communications. His research group has received several paper awards. His paper received the best paper award at the 2000 speech coding workshop and his students have received student paper awards at both the 2005 and 2006 International conference on Acoustics, Speech and Signal Processing conference as well as the best student paper award at NIPS 2006. A paper he co-authored with B. Song and R. Cruz received the 2008 Stephen O. Rice Prize Paper Award in the Field of Communications Systems and a paper he co-authored with S. Shivappa and M. Trivedi received the best paper award at AVSS 2008. He also received the graduate teaching award from the graduate students in the Electrical Engineering department at UCSD in 1998. He was elected to the fellow grade in 2000 for his contributions in high resolution spectral estimation. Dr. Rao has been a member of the Statistical Signal and Array Processing technical committee, the Signal Processing Theory and Methods technical committee as well as the Communications technical committee of the IEEE Signal Processing Society. He currently serves on the editorial board of the EURASIP Signal Processing Journal.