JUST-IN-TIME MULTIMODAL ASSOCIATION AND FUSION FROM HOME ENTERTAINMENT

Danil Korchagin[1], *Petr Motlicek*[1], *Stefan Duffner*[1], *and Hervé Bourlard*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland

ABSTRACT

In this paper, we describe a real-time multimodal analysis system with just-in-time multimodal association and fusion for a living room environment, where multiple people may enter, interact and leave the observable world with no constraints. It comprises detection and tracking of up to 4 faces, detection and localisation of verbal and paralinguistic events, their association and fusion. The system is designed to be used in open, unconstrained environments like in next generation video conferencing systems that automatically "orchestrate" the transmitted video streams to improve the overall experience of interaction between spatially separated families and friends. Performance levels achieved to date on hand-labelled dataset have shown sufficient reliability at the same time as fulfilling real-time processing requirements.

Index Terms — Multimodal signal processing, data analysis, association rules, sensor fusion

1. INTRODUCTION

The TA2 (Together Anywhere, Together Anytime) project [1] tries to understand how technology can help to nurture family-to-family relationships to overcome distance and time barriers. This is something that current technology does not address well: modern media and communications are designed for individuals, as phones, computers and electronic devices tend to be user centric and provide individual experiences. Existing multiparty conferencing solutions available on the market such as Microsoft RoundTable conferencing table [2] are not designed to be used in open, unconstrained environments.

In our work, we target the next generation of orchestrated video conferencing systems with spatially separated non-intrusive sensors. We are particularly interested in efficient mechanism of just-in-time multimodal cue association and fusion in open, unconstrained environments to be employed by a subsequent reasoning step. The reasoning produces then an orchestrated video chat [3] by choosing at each point in time the perspective that best represent the social interaction.

The fusion can be performed at different levels, based on type of input information available. It can be at sensor level, feature level, score level, rank level or decision level. First two levels can be considered as pre-classification category, while others can be considered as post-classification category [4]. The feature-level multimodal approach is normally done via transformation of the data in such a way that a correlation between the audio and a specific location in the video is found [5, 6]. In our study we concentrate mainly on score level fusion and propose a technique, which relies on information derived from spatially separated sensors. By placing the sensors at their individually optimal locations, we clearly obtain a better performance of low-level semantic information. This in turn results in good performance of the complete system. Other score-level multimodal techniques rely on estimation of the mutual information between the average acoustic energy and the pixel value [7], probability densities estimation [8] or a trained joint probability density function [9]. A subsequent reasoning step, which is not part of this paper, relies on decision-level rule-based fusion.

All described in the paper components were successfully integrated into a low delay analysis system. The detection and tracking of faces and the estimation of direction of sound (who spoke when?) are integrated into the main processing chain. This information is then used to associate events detected with the respective person IDs coming from the visual processing.

The association and fusion of acoustic and visual events is not a trivial task, because at each time instant there might be some events that are more reliable than others. The combined model has to be able to compute a confidence measure of the different modalities and weighs them accordingly. In addition, the sensors capturing the audio and video signals are spatially separated (as opposed to other systems, such as [10, 11], relying on collocated sensors).

In this context, TA2 presents several challenges: the results need to be computed in real-time with low affordable delay from spatially separated sensors in open, unconstrained environment. Furthermore, the results are supposed to be localised in the image space to allow for a dynamic and seamless orchestrated video chat.

2. A REAL-TIME ARCHITECTURE

The presented multimodal analysis system includes robust face tracking, far-field voice activity detection, estimation of direction of arrival, Automatic Speech Recognition (ASR) with keyword and proper name spotting. A face tracking algorithm has been developed to track a variable number of faces even when there is no face detection for a long period of time. Although the accuracy of far-field ASR is not yet good enough to be exploited for obtaining an accurate real-time transcription, it is already sufficient to augment the behaviour of an orchestration module. Words in the transcript are used to search for participants' proper names relevant to the group of people or keywords relevant to a given scenario. Furthermore, the orchestration (which is not part of the multimodal analysis system) will be able to reason and act upon these events together with other cues that could potentially come from a game engine, aesthetic or cinematic rules, making orchestrated video chat dynamic and seamless.

The system architecture is built around several modules (see Fig. 1) comprising a so-called Video Cue Detection Engine (VCDE) with a face detector, a multiple face tracker and multiple person identification; an Audio Cue Detection Engine (ACDE) with a direction of arrival estimator, a voice activity detector and a large vocabulary continuous speech recogniser; a Unified Cue Detection Engine (UCDE) performing association and fusion.

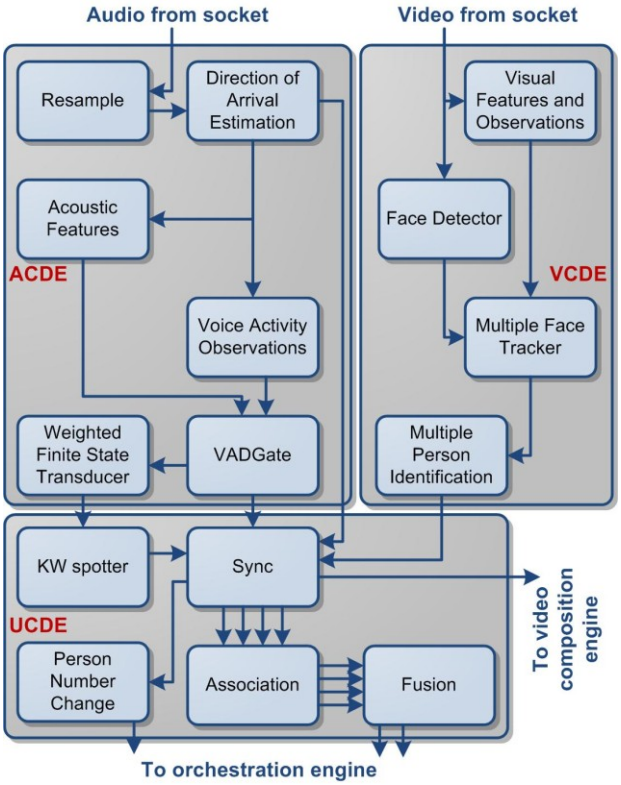

Fig. 1. Real-time architecture.

Both input to and output from the system are done via sockets. The core capture devices for the system are a FullHD video camera and an audio diamond array with four omnidirectional microphones [12]. The audiovisual streams are captured on an external server(s) and supplied to the analysis engine via sockets. The socket interface allows for a flexible software solution, though adds a latency of 12-20 ms for the audio stream and 30-300 ms for the video stream as the system requires uncompressed signals.

2.1. Framing and synchronisation

The multimodal processing operates in multi-framing mode with non-overlapping video frames at variable frame rate for video processing, overlapping audio frames of 16 ms in step of 10 ms for voice activity detection and ASR, and overlapping audio frames of 32 ms in step of 16 ms for direction of arrival estimation (see Fig. 2).

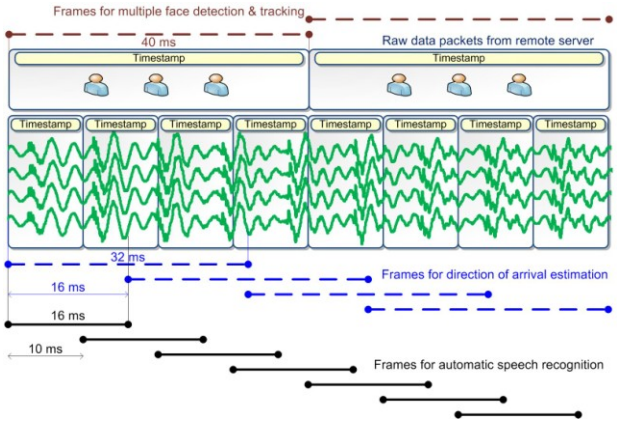

Fig. 2. Framing for online processing.

Video packets from the video grabber server are retrieved every 40-100 ms at a resolution of 640x360 pixels, while audio packets are retrieved every 10 ms and contain interleaved 4 channel PCM audio in 16-bit at 48 kHz. Each packet contains also unique 64-bit timestamp in microseconds for synchronisation between different remote modules.

2.2. Multiple face tracking

A multiple face tracking algorithm is automatically initialised and updated using outputs from a standard face detector [13]. The challenge for face tracking in this scenario is that face detections are not continuous and that the time between two successive detections can be very long (up to 30 s in our experiments). This is due to head poses that are difficult to detect by state of the art algorithms or partial occlusions caused by hands in front of the face. However, in the TA2 scenario it is necessary to know at each time instant where the people are in the video scene.

The solution employed in this work is based on a multi-target tracking algorithm using Markov Chain Monte Carlo (MCMC) sampling, similar to [14]. This is a Bayesian tracking framework using particles to approximate the current state distribution. At each time step, targets are added according to the output of the face detector, and targets are removed if there has not been any detection associated to a target for more than 10 s, or if the likelihood drops drastically.

The state space is the concatenation of the states of all visible faces, where the state of each single face is a rectangle described by the 2D position in the image plane, a scale factor and the eccentricity (height/width ratio).

The dynamic model is the product of the models of each visible face and a Markov Random Field that prevents targets becoming too close to each other. The state dynamics of each single face is described by a first-order autoregressive model for the position and a zeroth-order model for scale and eccentricity.

Finally, the observation likelihood is the product of the observation likelihoods of each visible face, which in turn is calculated using the Bhattacharyya distance between the HSV (Hue-Saturation-Value) colour histograms over three horizontal bands on the face region and the respective reference colour histograms which are initialised when the face is detected.

2.3. Multiple person identification

Whenever a tracker loses a target and reinitialises it later on, or a person leaves the visual scene and comes back later, the tracking algorithm tries to recognise the respective person in order to associate it to a previously tracked target. This is not done inside the tracking algorithm but on a higher level taking into account longer-term visual appearance observations. We have found that HSV colour histograms calculated on face and shirt regions yield a simple but effective measure of visual similarity. When identifying a "new" face, the current colour histograms are compared to the stored models of all previously seen people and if the similarity is above a certain threshold the corresponding ID is assigned, otherwise a new person model is created.

2.4. Direction of arrival estimation

Speaker localisation is performed by the direction of arrival module (Fig. 1). The algorithm is based on short-term clustering of generic sector-based activity measures [12, 15] in steps of 5°. It relies only on the geometry of the microphone array and does not depend on prior knowledge of the room dimensions. It can be effectively used to both detect and localise multiple sources in open, unconstrained environments.

2.5. Voice activity detection

Voice activity detection (VAD) covers both verbal and paralinguistic activity and is implemented as a gate. The gate segments the input stream in accordance to Boolean voice activity / silence information from a VAD algorithm based on silence models or trained multi-layer perceptrons (MLP) using traditional ASR features [16].

2.6. WFST decoder and keyword spotter

The ASR component is represented by the Weighed Finite State Transducer (WFST) based token passing decoder known as Juicer [16]. The output from the decoder is used to perform the spotting of proper names and keywords. More specifically, the spotting is performed based on the predefined list of participants and keywords relevant to the given scenario (e.g., orchestrated video chat).

2.7. Association and fusion

Due to the real-time requirements, the association and fusion of person IDs from the video identification with voice activity cannot be postponed until the voice activity is over. The fused events have to be published to the orchestrator within a few hundreds of milliseconds to keep the feeling of virtual presence and togetherness. The low delay temporal association and fusion scheme is depicted in Fig. 3.

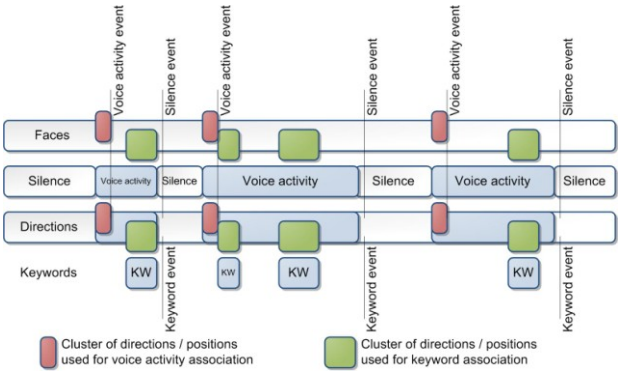

Fig. 3. Low delay association and fusion.

Since the position of people does not significantly change within a few hundred milliseconds, predictive temporal association can be used for video modality. Furthermore, audiovisual association is performed between acoustic short-term directional clusters and the positions of detected faces from the video modality. This involves a mapping estimation between microphone array coordinates (acoustic directional clusters w.r.t. the microphone array centre) and the coordinates of the image plane, which are defined by the field of view of the camera (Fig. 4).

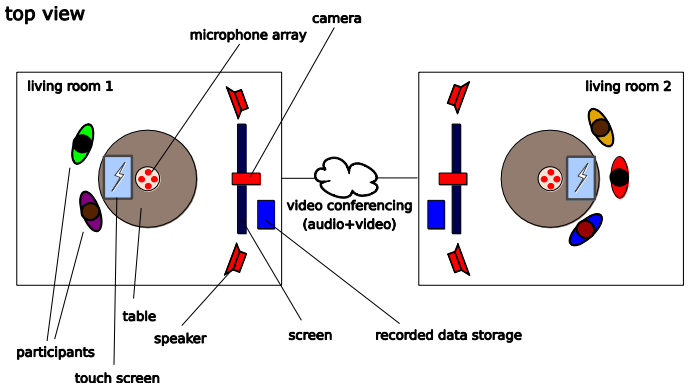

Fig. 4. TA2 setup, view from top [17].

Since the participants do not sit at predefined positions in the room, it can often cause ambiguities in the association and fusion. Clearly, the same acoustic short-term directional cluster can correspond to different positions in the image and vice-versa. Therefore, the location of a detected face within the image can be mapped to many different sound directions. However, since the participants are mainly located around a coffee table, such ambiguities occur rarely. Therefore, given the mean angle α of the directional cluster from the audio modality, a simplified association can be computed as:

$$\hat{i} = \arg\min_{i \in P} |x_i - x_{ma} - \gamma_i \sin\alpha|.$$

In the above formulation, P is the set of detected participants from the video modality, x_i is the horizontal position of the *i*-th person, x_{ma} is the horizontal position of the microphone array, and γ_i are calibration parameters.

3. EXPERIMENTAL RESULTS

The experiments were performed on real life hand-labelled datasets (3 h 50 min for Dataset 1 with enabled echo suppression [18]; 1 h 20 min for Dataset 2 [17] with disabled echo suppression, lower SNR and fewer frontal face views). The datasets follow the systematic description presented in [17] and contain recorded gaming sessions with enabled video chat of socially connected but spatially separated people. Each room was recorded and analysed separately and contained up to 4 people. The latency was measured on an Intel Core 2 CPU 6700 2.66GHz.

The achieved F-measure at different steps of processing on described datasets is summarised in Table 1. The F-measure is defined as the harmonic mean of precision and recall values:

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Precision is defined as the number of true positive test events (test events correctly detected as belonging to the positive class) divided by the total number of test events detected as belonging to the positive class (the sum of true positive and false positive test segments). Recall is defined as the number of true positives test events divided by the total number of test events that actually belongs to the positive class (the sum of true positive and false negative test events). The annotated voice activity events from Dataset 2 are illustrated in Fig. 5.

Table 1. F-measure at different steps of processing

Algorithm	F-measure	
	Dataset 1	Dataset 2
Face detection [13]	73.7%	67.1%
Local far-field VAD	81.9%	69.1%
Acoustic localisation	93.1%	89.4%
Multiple face tracking	89.1%	89.4%
Person localisation, based on fusion of AV information	89.2%	83.7%
Local far-field VAD, based on fusion of AV information	80.2%	69.0%

The first row of the Table 1 shows the F-measure of a standard face detector [13] applied on single frames of the video stream. It represents mean value over all people. The 4th row shows the results of the face tracking algorithm, which improves the overall accuracy of the video processing.

The second and third rows show the F-measure on the output of the local far-field voice activity detection and acoustic localisation (6000+ observations). Since Dataset 1 is echo-cancelled and less noisy, we were able to detect local voice activity with much higher precision/recall than for Dataset 2, in which echo from remote location negatively affects the precision level of local far-field VAD.

Far-field voice activity detection based on fusion of AV information is given in row 6. One can see that the performance can vary in comparison to row 2 due to assigning the voice activity to video tracked person. Row 5 expresses a person identity association of detected voice activity, i.e. how well we can assign previously detected voice activity to a local person based on AV information.

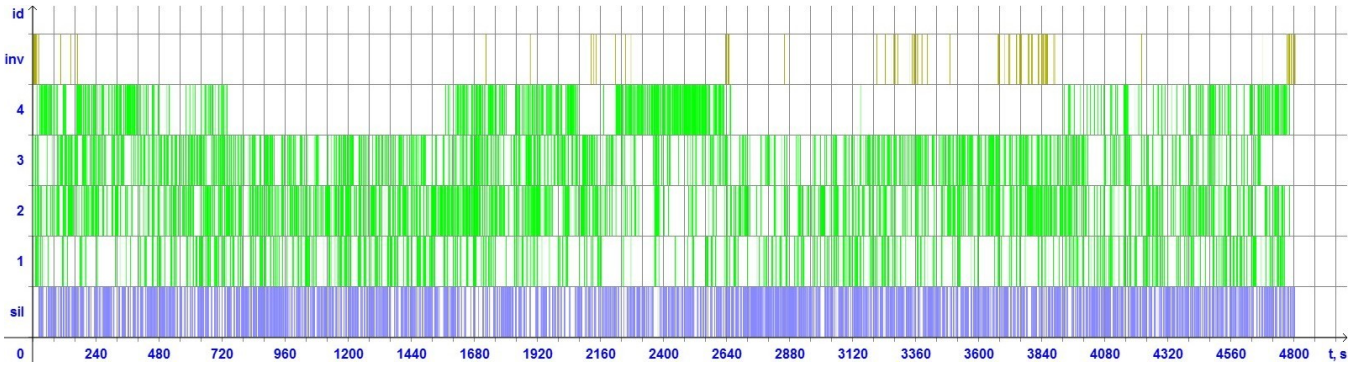

Fig. 5. Annotated voice activities over time for room 1 of Dataset 2 [17]. The first row (inv) shows voice activity of persons not visible from the camera, the four following rows show the voice activity of the four different persons visible in the video (id 1 to 4), and the last row shows silence (sil). One can see that there are a lot of short utterances, and speakers change quite frequently.

4. CONCLUSION

We have developed a framework for just-in-time multimodal association and fusion for open, unconstrained environments with spatially separated multimodal sensors. Performance levels achieved to date on hand-labelled echo-cancelled dataset have shown sufficient reliability at the same time as fulfilling real-time processing requirements with latency within 200-300 ms. The achieved results are promising for the further development of the platform in several directions such as improvement of performance, reduction of the latency and integration of elevation component into fusion to allow analysis of another layer of people, who could be standing up behind.

5. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme ICT Integrating Project “Together Anywhere, Together Anytime” (TA2, FP7/2007-2013) under grant agreement no. ICT-2007-214793. We are grateful to Fraunhofer IIS for provision of the real life echo-cancelled dataset, Philip N. Garner and Jean-Marc Odobez for their valuable help at various stages of this work.

6. REFERENCES

[1] Integrating project within the European research programme 7, "Together anywhere, together anytime", http://www.ta2-project.eu, 2008.

[2] Microsoft, "Microsoft RoundTable conferencing table", http://www.microsoft.com/uc/products/roundtable.mspx, 2007.

[3] M. Falelakis, et al., "Reasoning for Video-mediated Group Communication", Proc. IEEE International Conference on Multimedia & Expo (ICME), Barcelona, Spain, 2011.

[4] C. Sanderson and K. K. Paliwal, "Information fusion and person verification using speech and face information", Idiap Research Report IDIAP-RR 02-33, 2002.

[5] M. Slaney and M. Covell, "Facesync: a linear operator for measuring synchronization of video facial images and audio tracks", Proc. of Neural Information Processing Systems, pp. 814-820, 2000.

[6] G. Monaci, O. D. Escoda, and P. Vandergheynst, "Analysis of multimodal sequences using geometric video representations", in Signal Processing, vol. 86, pp. 3534-3548, 2006.

[7] J. Hershey and J. Movellan, "Audio vision: using audio-visual synchrony to locate sounds", Proc. of Neural Information Processing Systems, pp. 813-819, 1999.

[8] H. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: an empirical study", Proc. of CIVR, Urbana-Champaign, USA, 2003.

[9] M. Gurban, and J. Thiran, "Multimodal speaker localization in a probabilistic framework", Proc. of EUSIPCO, Florence, Italy, 2006.

[10] D. Bohus and E. Horvitz, "Dialog in the open world: platform and applications", Proc. of ICMI, Cambridge, USA, 2009.

[11] K. Otsuka, et al., "A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization", Proc. of ICMI, Chania, Greece, 2008.

[12] D. Korchagin, P.N. Garner, and P. Motlicek, "Hands free audio analysis from home entertainment", Proc. of Interspeech, Makuhari, Japan, 2010.

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", Proc. of CVPR, Hawaii, USA, 2001.

[14] Z. Khan, "MCMC-based particle filtering for tracking a variable number of interacting targets", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 27, pp. 1805-1918, 2005.

[15] G. Lathoud and I. A. McCowan, "A sector-based approach for localization of multiple speakers with microphone arrays", Proc. of Statistical and Perceptual Audio Processing, Jeju, Korea, 2004.

[16] P. N. Garner, et al., "Real-time ASR from meetings", Proc. of Interspeech, pp. 2119-2122, Brighton, UK, 2009.

[17] S. Duffner, P. Motlicek, and D. Korchagin, "The TA2 database: a multi-modal database from home entertainment", Proc. of Signal Acquisition and Processing, Singapore, 2011.

[18] F. Kuech, et al., "Acoustic echo suppression based on separation of stationary and non-stationary echo components", Proc. of Acoustic Echo and Noise Control, Seattle, USA, 2008.