Multimodal Cue Detection Engine for Orchestrated Entertainment

Danil Korchagin, Stefan Duffner, Petr Motlicek, and Carl Scheffler

Idiap Research Institute, Martigny, Switzerland

Abstract. In this paper, we describe a low delay real-time multimodal cue detection engine for a living room environment. The system is designed to be used in open, unconstrained environments to allow multiple people to enter, interact and leave the observable world with no constraints. It comprises detection and tracking of up to 4 faces, estimation of head poses and visual focus of attention, detection and localisation of verbal and paralinguistic events, their association and fusion. The system is designed as a flexible component to be used in conjunction with an orchestrated video conferencing system to improve the overall experience of interaction between spatially separated families and friends. Reduced latency levels achieved to date have shown improved responsiveness of the system.

Keywords: Multimodal signal processing, data analysis, 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 previous work [3], we have developed a framework for just-in-time multimodal association and fusion for open, unconstrained environments with spatially separated multimodal sensors. It relies on score-level information fusion 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. Performance levels achieved 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. In current work we evolve the previous system towards better responsiveness of the system and integration of additional components, which have been identified as important for the extraction of additional semantic cues to be used by an orchestration engine [4]. The orchestration engine produces then an orchestrated video chat by choosing at each point in time the

perspective that best represents the social interaction based on decision-level rule-based fusion.

Fig. 1. Illustration of a family environment setup.

In this context, TA2 presents several challenges: the results need to be computed in real-time with low affordable delay from spatially separated sensors (as opposed to other systems, such as [5, 6, 7], relying on collocated 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 cue detection engine includes a face detector, a multiple face tracker, multiple person identification, head pose and visual focus of attention estimation, an audio real-time framework with spatial localisation, a large vocabulary continuous speech recognizer and keyword spotter, multimodal association and fusion (see Fig. 2). 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 Automatic Speech Recognition (ASR) is not yet good enough to be exploited for obtaining an accurate real-time transcription, it is 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 cue detection engine) 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.

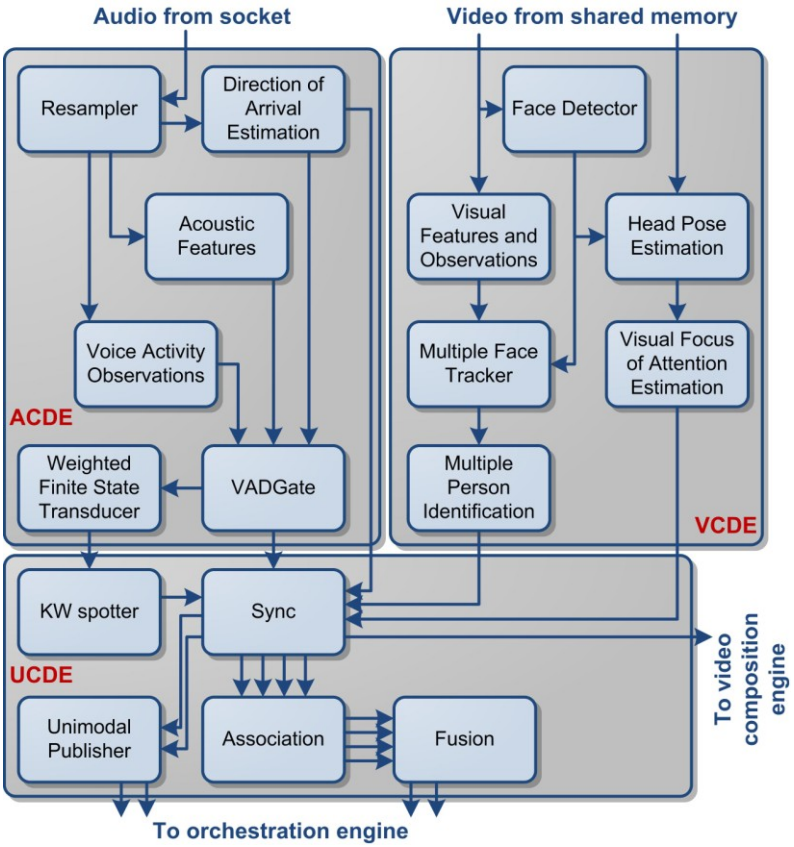

Fig. 2. The system architecture is built around several modules comprising a so-called Video Cue Detection Engine (VCDE) with a face detector, a multiple face tracker, multiple person identification, head pose and visual focus of attention estimation; an Audio Cue Detection Engine (ACDE) with a direction of arrival estimator, a voice activity detector and a large vocabulary continuous speech recogniser; a Unified Cue Detection Engine (UCDE) with association, fusion and transmission of the results to external components (orchestration engine, video composition engine).

The audio input to the multimodal cue detection engine and the semantic output from it are implemented via sockets, while the video stream is transferred via shared memory. The core capture devices for the system are a Full HD video camera and an audio diamond array with four omnidirectional microphones [8]. Video frames from the shared memory of the video grabber server are retrieved every 40 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.

The multimodal processing operates in multi-framing mode with non-overlapping video frames, 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.

2.1 Multiple Face Tracking

A multiple face tracking algorithm is automatically initialised and updated using outputs from a standard face detector [9]. 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 (see Fig. 3). However, in the TA2 scenario it is necessary to know at each time instant where the people are in the video scene.

Fig. 3. An example of difficult to detect head poses and partial occlusions [10].

The solution employed in this work is based on a multi-target tracking algorithm using Markov Chain Monte Carlo (MCMC) sampling, similar to [11]. This is a Bayesian tracking framework using particles to approximate the current state distribution of all visible targets. At each time step, targets are added and removed using the output of an additional probabilistic framework that takes into account the output of the face detector as well as long-term observations from the tracker and image [12].

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 are 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.

2.2 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. Each person's appearance is modelled by three sets of HSV colour histograms calculated on face and shirt regions. Using multiple histograms per person copes for different appearances due to changes in body pose. However, only the most similar histogram of a person is used and updated at each time.

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.

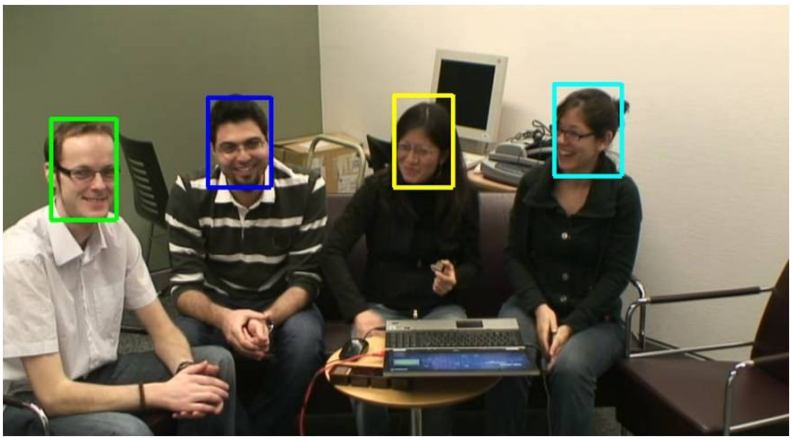

Fig. 4. Consistent person identification within the session (here indicated by different colours) is an important requirement to the multimodal cue detection engine.

2.3 Head Pose Estimation

Based on the output of the face tracker, the head pose (i.e. rotation in 3 dimensions) of an individual is estimated. The purpose of computing head pose is the estimation of a person's visual focus of attention (see section 2.4).

Head pose is computed using visual features derived from the 2-dimensional image of a tracked person's head. The features used here are gradient histograms [13] and colour segmentation histograms. Colour segmentation is done by classifying each pixel around the head as either skin, hair, clothing or background based on colour models that are adapted to each individual being tracked [14].

To compensate for the variability in the output of the face tracker, the 2-dimensional face location is re-estimated by the head pose tracker. This serves to normalise the bounding box around the face as well as possible, while simultaneously using the visual features mentioned above to estimate pose. This joint estimation of head location and pose improves the overall pose accuracy [15].

2.4 Visual Focus of Attention

Given the estimated belief (probability distribution) over head pose, the visual focus of attention target is estimated. In the context of this work, the following targets are of interest: the video conferencing screen, the touch sensitive table, and any other person in the room.

The range of angles that correspond to each target is modelled using a Gaussian likelihood. This likelihood is derived from the known spatial locations of the targets within the conference room. The posterior belief over each target is computed with Bayes' rule using the method given in [16].

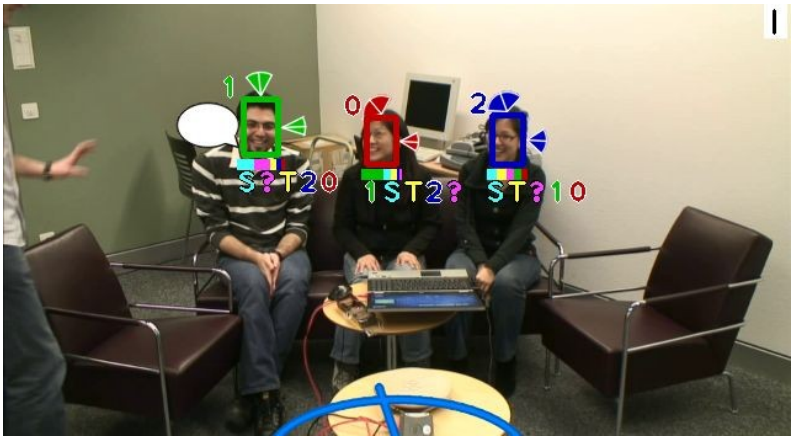

Fig. 5. Multimodal cue visualisation. For each person, it shows its ID (at the top-left of the face bounding box), its head orientation estimation, i.e. pan and tilt, with a variance indication (on the top and right side of the box), and the estimated distribution over targets where the person is looking at (at the bottom of the box), where the left-most target is the most likely one. The letter "S" means "screen", "T" means "table", "?" means "unknown", and the numbers correspond to the IDs of the other persons. The blue line in the bottom of the image indicates the estimated direction of arrival of sound. The speech bubble indicates that a person is speaking, and the output of the keyword spotting is shown in the top-right of the image, here the word "I".

2.5 Direction of Arrival Estimation

Speaker localisation is performed by the direction of arrival module (Fig. 2). The algorithm is based on spatio-temporal fingerprint processing [17] in steps of 6°, which represents a computationally efficient solution with low algorithmic delay compared to short-term clustering of generic sector-based activity measures [8, 18] used in our previous study [3]. 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.

2.6 Voice Activity Detection

Voice activity detection (VAD) covers both verbal and paralinguistic activities and is implemented as a gate. The gate segments the input stream in accordance to directional and voice activity / silence information from an algorithm based on silence models or trained multi-layer perceptrons (MLP) using traditional ASR features [19]. The association and fusion [3] of the detected voice activity events with person IDs from the video-based identification are performed by the time voice activity is confirmed and the corresponding audio-based directional cluster is estimated.

2.7 Keyword spotting

The ASR component is represented by the Weighed Finite State Transducer (WFST) based token passing decoder known as Juicer [19]. The output from the decoder is used to perform the spotting, association and fusion [3] of proper names and keywords with person IDs from the video identification taking into account the estimated audio-based directional cluster for the corresponding time interval. 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).

3 Improvements and Results

During subjective evaluations of our previous version of multimodal cue detection engine, several bottlenecks have been experienced. To overcome these bottlenecks, several architectural and algorithmic changes have been applied and presented in this paper.

First of all, while the socket interface was allowing for a flexible software solution, the experienced latency for uncompressed video signal transmission from remote video grabber was resulting in additional latency of 30-300 ms. This clearly noticeable lag was successfully removed by switching to a shared memory interface for video input stream. While a shared memory interface could be potentially used for audio input stream as well, experienced latency of 12-20 ms for the audio transmission is on an acceptable level.

To reduce the latency of audio processing we have decided to reduce the algorithmic delays of both direction of arrival estimation and voice activity detection. The algorithmic latency of both components has been reduced from 200 ms down to 128 ms. This is due to the replacement of the previous implementation based on a short-term clustering approach by the computationally more efficient spatio-temporal fingerprints processing and the reduction of corresponding temporal filters.

Exact clock synchronisation between separated audio and video grabbers was seen as another source of potential problems and during subjective evaluations we have found that the use of local timestamps results in more consistent multimodal association and fusion. Moreover, since the position of people does not significantly change within a few hundred milliseconds, predictive temporal association was finally employed within the system to further remove possible lags during the capturing of the video stream by hardware and video grabber.

We have found that it is beneficial to have acoustic tracking of the active acoustic sources as an additional input to the voice activity detection gate to properly treat barge-in events, which were not always detected in a former system.

Since the participants do not sit at predefined positions in the room, theoretically it can cause ambiguities in the association and fusion. Clearly, the same acoustic directional cluster can correspond to different positions in the image and vice-versa. However, since the participants are mainly located around a coffee table, such ambiguities occur rarely during evaluations.

Finally, head pose and visual focus of attention estimation have been identified as important semantic cues for the orchestration engine and have been successfully integrated into the multimodal cue detection engine. Head pose estimation is to be used for better selection of frontal/side views with respect to aesthetic and cinematic rules, while visual focus of attention can be beneficial for better modelling of social interactions (e.g. predictive turn estimation during grant-floor moments) and can have a direct impact on temporal filters within the aesthetic and cinematic rules.

Objective evaluations of involved components were performed, and their results can be found in [3, 12, 14, 17]. The corresponding annotated dataset has been made publically available [10]. The algorithmic latency within the multimodal cue detection engine stays within 130 ms, except for proper name and keywords spotting, which are transmitted by the end of acoustically separated utterances.

4 Conclusion

We have developed a low delay real-time multimodal cue detection engine for open, unconstrained environments with spatially separated multimodal sensors. We have described applied architectural and algorithmic changes to reduce an overall latency down to 130 ms and fulfil real-time processing requirements. The achieved results are promising for future wider evaluations and further development of the platform in several directions such as improvement of performance, reduction of the latency, and integration of additional components allowing richer multimodal cues.

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 Philip N. Garner and Jean-Marc Odobez for their valuable help at various stages of this work.

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. Korchagin, D., Motlicek, P., Duffner, S. and Bourlard, H.: Just-in-time multimodal association and fusion from home entertainment. In: Proc. IEEE International Conference on Multimedia & Expo (ICME), Barcelona, Spain (2011)
4. Falelakis, M. et al.: Reasoning for video-mediated group communication. In: Proc. IEEE International Conference on Multimedia & Expo (ICME), Barcelona, Spain (2011)
5. Bohus, D. and Horvitz, E.: Dialog in the open world: platform and applications. In: Proc. of ICMI, Cambridge, USA (2009)
6. Otsuka, K. et al.: A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization. In: Proc. of ICMI, Chania, Greece (2008)
7. Bernardin, K., Stiefelhagen, R.: Audio-visual multi-person tracking and identification for smart environments. In: Proc. of ACM Multimedia (2007)
8. Korchagin, D., Garner, P.N., and Motlicek, P.: Hands free audio analysis from home entertainment. In: Proc. of Interspeech, Makuhari, Japan (2010)
9. Viola, P. and Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. of CVPR, Hawaii, USA (2001)
10. Duffner, S., Motlicek, P., and Korchagin, D.: The TA2 database: a multi-modal database from home entertainment. In: Proc. of Signal Acquisition and Processing, Singapore (2011)
11. Khan, Z.: MCMC-based particle filtering for tracking a variable number of interacting targets. In: IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 27, pp. 1805–1918 (2005)
12. Duffner, S., Odobez, J.-M.: Exploiting long-term observations for track creation and deletion in online multi-face tracking. In: Proc. IEEE Conference on Automatic Face & Gesture Recognition (2011)
13. Dalal, N. and Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (2005)
14. Scheffler, C. and Odobez, J.-M.: Joint adaptive colour modelling and skin, hair and clothing segmentation using coherent probabilistic index maps. In: Proc. of BMVC (2011)
15. Ba, S.O. and Odobez, J.-M.: A probabilistic framework for joint head tracking and pose estimation. In: Proc. of the International Conference on Pattern Recognition (2004)
16. Ba, S.O. and Odobez, J.-M.: Recognizing visual focus of attention from head pose in natural meetings. In: IEEE Transactions on System, Man and Cybernetics, 39(1):16–33 (2009)
17. Korchagin, D.: Audio spatio-temporal fingerprints for cloudless real-time hands-free diarization on mobile devices. In: Proc. of the 3rd Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA), pp. 25–30, Edinburgh, UK (2011)
18. Lathoud, G. and McCowan, I. A.: A sector-based approach for localization of multiple speakers with microphone arrays. In: Proc. of SAPA, Jeju, Korea (2004)
19. Garner, P. N., et al.: Real-time ASR from meetings. In: Proc. of Interspeech, pp. 2119–2122, Brighton, UK (2009)