

Processing Image and Audio Information for Recognising Discourse Participation Status through Features of Face and Voice

Nick Campbell¹, Damien Douxchamps²

¹ NiCT/ATR-SLC

National Institute of Information and Communications Technology
Keihanna Science City, Kyoto 619-0288, Japan

²Image Processing Laboratory,

Nara Institute for Science and Technology,
Nara 630-0192, Japan

nick@nict.go.jp, ddouxcha@is.naist.jp

Abstract

This paper describes a system based on a 360-degree camera with a single microphone that detects speech activity in a round-table context for the purpose of estimating discourse participation status information for each member present. We have obtained 97% accuracy in detecting participants and have shown that the use of non-verbal and backchannel speech information is a useful indicator of participant status in a discourse.

Index Terms: round-table meetings, image processing, non-verbal behaviour, speech activity, discourse management

1. Introduction

There has recently been considerable research activity into the processing of participant status in meetings and round-table discussions [1, 2, 3, 4]. This paper presents a novel approach for the detection of participant status in such situations, based on the use of a 360-degree camera and a single low-quality microphone to detect and analyse types of speech activity from the members. The speech is then classified into verbal or non-verbal variants and a decision is made about the speakers underlying intention for the utterance. Intentions are currently limited to the following classes: (a) provision of information or opinion, (b) brief comment, (c) laugh, (d) backchannel, (e) off-record personal communication. The system uses very low-level primitives such as head detection, body estimation, movement detection, and sound detection, and uses the combination and timing of activities in each to estimate participant status.

2. Image processing

Visual clues of the behaviour of discourse participants are extracted from the streaming video image by combining standard tools to form a more specialized video processing chain. Much of the processing is aimed towards a proper face detection since the face is a human feature that is relatively easy to detect and contains a lot of information concerning the behaviour of the person. Detecting hands is also an option but these are more difficult to track as their shape can vary greatly and they also move much faster. This in turn requires a higher video framerate which weighs heavily on the processing speed. Our process for detecting and characterizing faces is thus as follows:

First, the video signal from a digital camera is decoded from a raw Bayer format to a full RGBI image. The algorithm used

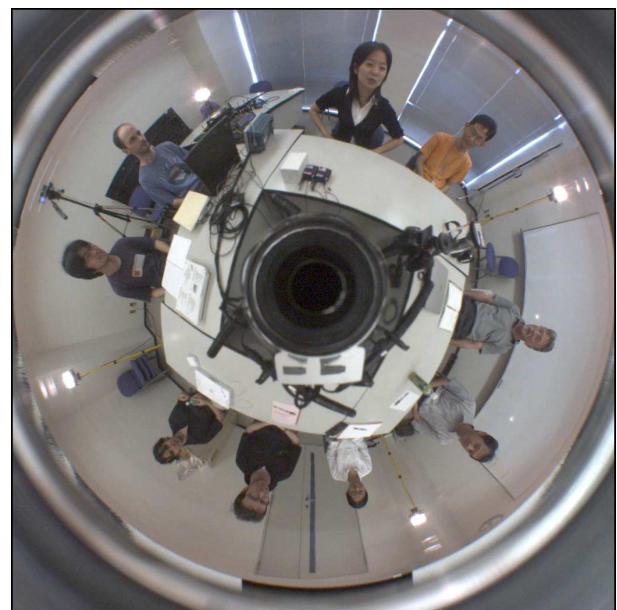


Figure 1: Circular 360 degrees image captured by the camera



Figure 2: Rectified 360 degrees image

is the 'Edge Sense II' presented in [5]. This algorithm provides good quality output while still being able to run at a reasonable speed. Other algorithms have been used [6] [7] but did not provide a significant advantage while being considerably slower. At this point the image still consists of a circular band (fig.1) and must be rectified before the face detection. To this end a simple linear resampling is performed. The resulting rectified image (fig.2) is now ready to be used for the face detection.



Figure 3: Typical output from the program showing two 180 degrees sections on top of each other. Detected faces are shown with a white square, tracked faces with a black square.

2.1. Face Detection And Tracking

Face detection can be performed in a number of ways. The first technique that we tried was based on background and colour [8] segmentation and failed to provide satisfactory results due to illumination changes and colour restrictions. A better approach is to use the Viola-Jones face detection [9] [10] which is based on pattern matching. One drawback of this approach is that the algorithm must be trained on a large number of images. However, standard packages such as OpenCV provide example training data (in the form of Haar cascades) that we found to be very effective to detect the two patterns that we are most interested in: profile faces and frontal faces. In fact, using the Viola-Jones detection alone more than 60% of faces can be found in our round-table data. Adding a simple skin color check and a face size check on the detected regions limits false positives to almost zero. Note that it is also necessary to remove duplicate (overlapping) faces since Viola-Jones can detect two instances of a single face. Also, mathematical morphology was used to limit the effect of color noise in the video during the skin color check.

The Viola-Jones face detection is strictly frame-based. The lack of time integration means that the detection can oscillate even with small image variations: a face can be detected in frame t , disappear in frame $t + 1$ and appear again in frame $t + 2$. To avoid these instabilities a method of tracking the faces was introduced. Faces detected on a previous image will be matched with faces found on the current image. If no match is found then the old face will be tracked in the new image in order to cover the gap in detection. The tracking is performed using a classic block-matching algorithm (BMA).

The tracking can drift in time so it is necessary to limit it with some safeguards. The first one consists in limiting the time during which this gap-bridging tracking will be performed. Given the good quality of the faces detected by Viola-Jones we can allow a long tracking time of around 30 seconds. The second limitation consists in verifying that the tracked face still contains a minimal amount of skin-coloured area (20%). Thirdly, the image difference between the old and tracked faces should be limited. Finally, the amount of face motion is also limited by the size of the search zone of the BMA.

At this point we have not yet included any situation-specific verifications that may help to filter out the last outlying faces. To remain as generic as possible we only include one: if two faces are overlapping vertically (i.e., if they belong to the same image column) then only the highest face is kept. This is a small restriction that remains valid for most 'meeting' situations.

The resulting face detection processing sequence is able to detect more than 97% of faces during a one hour recording of a meeting at 10 frames per second (44000 frames total). At the same time the amount of false positives is very low: less than 0.01% or 10 occurrences for the whole sequence. A typical output image of our program is shown in fig.3.

2.2. Motion And Activity Estimation

Once faces are properly detected a number of measurements are performed to identify the face position, its motion and its area. The motion estimation cannot be performed on the positions of the detected faces because they are too unstable; parasitic motions of ± 5 pixels are not uncommon with the Viola-Jones detection. The motion estimation is therefore performed using subpixel block matching (BMA) on the image content.

A measure of the 'activity' of the person is provided separately for the face and body (the body being defined as the area below the face). This activity is computed as the mean pixel-based difference between the previous and the current image.

The graphs on fig.4 show a small section of five minutes of head and body measures (such as activity and motion) for the nine persons attending the meeting. These graphs show that the vertical and horizontal motion estimation of the face is able to resolve small details. For example persons mimicking a 'yes' or 'no' head movement are visible as small sinewave bursts. Activity measures also correlate well with more the global movements of a person. These measurements are now being correlated more systematically with manually labeled audiovisual data.

3. Audio Processing

The audio signal is captured by a single microphone and processed in parallel with the image. Level detection is used to determine the presence or absence of speech after adjusting for the background noise of the room environment. When speech is detected, it is compared with a duration threshold. Anything longer than 2 seconds is considered as containing information or opinion, anything less as a form of backchannel utterance. The former is tagged but not processed further, the latter is used to provide cues to the participation status of the individual members. The threshold was determined from the duration of laughs occurring in conversational speech: median duration 1.228 seconds, 75th quantile 1.931 seconds, from 3,393 occurrences.

3.1. Overlapping speech

We have tested the above equipment in three types of interactive speech situation. Our original goal was for use in a meetings environment, where between four and a dozen members would be seated around a table in a quiet office environment (see figures above). We extended this to a less formal meetings environment, where two or three friends or family members were engaged in conversation around a small table, such as over a meal or coffee. The third context included more people in a very relaxed party situation where alcohol and light snacks were consumed and the conversation was lively.

These discourse contexts can be differentiated primarily by the amount of overlapping speech that occurs in each. For the first, we have shown in previous work [4] that overlapping speech occurred less than 15% of the time. For the second, we found an average of 30% overlapping speech, but noted particular characteristics which we termed "Active Listening" [11] to be discussed further in the following subsection. For the third,

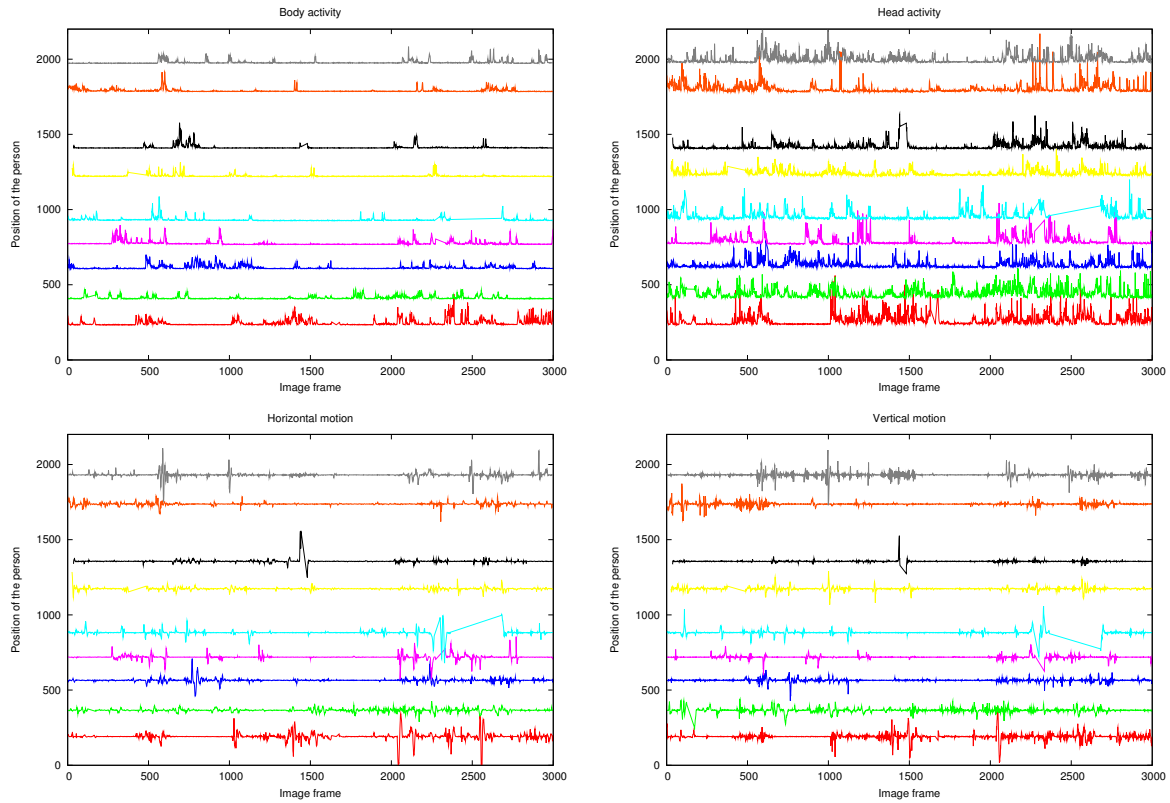


Figure 4: The body and head activity (top row) and the head horizontal and vertical movement (bottom row) of the 9 participants

the amount of overlapping speech was considerable, and no automatic processing was possible. With so many people talking at the same time, the social structure of the meeting collapses and it can be better considered as a cluster of small meetings, with each requiring its own sub-processing unit.

3.2. Backchannel utterances

In the formal situation of an office meeting, silence is the norm, with only one speaker active at any time. However, there are frequent occasions where laughter breaks out, or other noises of assent or dissent. Occasionally the structure of the meeting breaks down and people lapse into a private-conversation mode.

We have noted that in these more intimate conversations a form of active listening occurs, where the ‘listener’ produces overlapping speech. Table 1 gives details of speech activity timing, per speaker, over a set of 100 30-minute conversations. Data were calculated from the time-aligned manual transcriptions. The median speaking time is approximately 18 minutes per speaker. There is approximately 3 minutes when no-one is speaking at all (10% of the total time) and 7 minutes (i.e., more than 20% of the conversation) when both speakers are speaking at once. Since time of non-overlapping speech is approximately 14 minutes per speaker, we can conclude that people overlap their speech, or talk simultaneously, one third of the time. These short feedback utterances provide essential information about participation status and listener attention.

3.3. Information or Opinion

It is not our goal to recognise the content of the discussions. Any utterance of more than one second in length is annotated as

Table 1: Showing quantiles of speech activity time per speaker. ‘Silence’ is when neither is speaking, ‘overlap’ when both are speaking at the same time. ‘silX’ (X = A or B) shows the time each speaker individually was quiet. ‘SoloX’ shows the total duration of non-overlapping speech per speaker, and ‘talkX’ the total overall speech time including overlaps. ‘Duration’ shows timing statistics for the entire conversation (assumed to be 30 minutes by default). All times are shown in minutes.

	min	25%	median	75%	max
silence	0.99	2.08	2.85	3.81	7.03
silA	6.73	10.68	14.02	16.91	22.46
silB	5.72	13.09	14.68	17.68	21.58
soloA	4.14	9.51	11.66	14.68	18.17
soloB	4.55	8.39	10.64	13.32	18.90
overlap	2.66	5.53	7.01	9.04	12.80
talkA	10.80	16.04	18.75	22.44	28.52
talkB	12.20	15.66	17.93	20.15	27.15
duration	28.57	32.00	32.93	33.96	37.98

such and left for future processing. However, we can estimate the opinions of the speakers from their choice of short utterance.

From an analysis of many transcribed conversations we found almost half of the utterances to be non-verbal; i.e., they could not be adequately understood from their text alone. Table 2, from [11], provides detailed figures for one speaker. Very few of these utterance types can be found as an entry in a standard language dictionary, yet it was confirmed experimentally that

Table 2: Counts of non-verbal utterances noted in the transcriptions for one female speaker in the conversation corpus. Utterances labelled ‘non-lexical’ consist mainly of laughs, grunts, and sound sequence combinations not typically found in a language dictionary. They also include common words such as ‘yeah’ ‘oh’, ‘uhuh’, etc.

total number of utterances transcribed	148772
number of unique ‘lexical’ utterances	75242
number of ‘non-lexical’ utterances	73480
number of ‘non-lexical’ utterance types	4492
proportion of ‘non-lexical’ utterances	49.4%

the intended meanings of many of these conversational noises can be perceived consistently by listeners even when presented in isolation without any discourse context information.

A speech recognition engine has been trained to recognise these non-verbal speech segments, with the intervening sections of normal speech treated as OOV (out of vocabulary) elements. The vocabulary size is currently 250, determined from the counts of repeated tokens.. Once a non-verbal element has been recognised, its prosodic characteristics are then compared against a database of features recorded for similar elements that have already been annotated for affect-related characteristics such as degree of familiarity, amount of interest, speaker activity, etc., and a prediction made about the characteristics of the current speech segment. These are then compared with the activity measure from the image processing module.

4. Integration of Image & Audio

Bodily movement is closely integrated with speech [12]. By watching who is moving when a sound is detected, we can use the movement to provide further information about the speech. At the lowest level, this provides speaker identification, but from the synchrony of their movements, and from their timing and duration, we can also estimate the function of each utterance. The amount of movement is a form of prosody that adds robustness to the estimation produced from the voice alone. We are currently training SVM and GMM models to predict human annotated labels from the above data. Recognition rates for the audio alone are 64%, and for the combined audio and image data 68% on open testing. This training continues as future work.

5. Conclusions

This paper has presented details of a system for the detection of participant activity in a round-table or meetings situation. It consists of a small camera and microphone arrangement that can be placed discretely on the tabletop for the streaming of image and audio data. The image is processed to obtain an estimate of speakers present and to produce measures of their various movements, which are then aligned with the speech sounds to identify the speaker and produce an estimate of the intention underlying each utterance. Utterances are broadly classified into verbal versus non-verbal classes, with the non-verbal tokens further analysed to distinguish various types of participation cues. Although a reduced version of this system works in real-time, the current system is very slow (factor 10 from real time). From now the image-processing research will now be

aimed at accelerating the processing and using the input from speech research to fine tune the algorithms.

6. Acknowledgements

The first author is supported by NiCT, the National Institute for Communications and Information Technology. This work was partially funded under the SCOPE initiative. Both are under the Japanese Ministry of Internal Affairs and Communications,

7. References

- [1] S. Burger, V. MacLaren, and H. Yu, “The ISL meeting corpus: The impact of meeting type on speech style”, in Proc. International Conference on Spoken Language Processing (ICSLP), Denver, Sept. 2002.
- [2] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, “The ICSI meeting corpus”, in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Hong-Kong, Apr. 2003.
- [3] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, “Automatic analysis of multimodal group actions in meetings”, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 27, no. 3, pp. 305-317, Mar. 2005.
- [4] W. N. Campbell, “A multimedia database of meetings and informal interactions for tracking participant involvement and discourse flow”, in Proc LREC 2006, Lisbon.
- [5] T. Chen, “A Study of Spatial Color Interpolation Algorithms for Single-Detector Digital Cameras”, <http://www-ise.stanford.edu/tingchen/>.
- [6] K. Hirakawa and T. W. Parks, “Adaptive Homogeneity-Directed Demosaicing Algorithm”, IEEE Trans. on Image Processing, vol. 14, no. 3, pp. 360-369, Mar. 2005.
- [7] E. Chang, S. Cheung and D. Pan, “Color filter array recovery using a threshold-based variable number of gradients”, in Proc. of the SPIE Conference, vol. 3650, pp. 36-43, Mar. 1999.
- [8] R.L. Hsu and M. Abdel-Mottaleb, “Face Detection in Color Images”, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 696-706, May 2002.
- [9] P. Viola and M. Jones, “Rapid Object Detection Using a Boosted Cascade of Simple Features”, in Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 511-518, Dec. 2001.
- [10] P. Viola and M. Jones, “Robust Real-Time Face Detection”, International Journal of Computer Vision, vol. 57, no. 2, pp. 137-154, May 2004.
- [11] W. N. Campbell, “Expressive Speech Processing and Prosody Engineering”, in *New Trends in Speech Based Interactive Systems*. Chen, F., & Jokinen, K., Springer, in (in Press).
- [12] W. S. Condon, “Communication: Rhythm and Structure. Rhythm in Psychological, Linguistic and Musical Processes”, J. R. Evans and M. Clynes. Springfield, Illinois, Charles C Thomas Publisher: 55-78. 1986.