"The Listening Eye": a non-verbal sensing device for interactive systems

Nick CAMPBELL [1] and Damien DOUXCHAMPS [2]

[1] National Institute of Information and Communications Technology, & ATR-SLC, Keihanna Science City, Kyoto 619-0288, Japan
[2]Image Processing Laboratory, Nara Institute for Science and Technology, Nara 630-0192, Japan
nick@nict.go.jp, ddouxcha@is.naist.jp

1 Brief description of the demonstration set-up

This demonstration presents a non-verbal sensor device which combines visual and audio information to provide an estimate of user behaviour for interactive dialogue systems such as voice-activated web-based information provision or humanoid robots. The setup consists of a notebook computer and a camera with a 360-degree lens. The system first detects potential partners in the vicinity of the computer, using peripheral vision and distinguishing between people who might be near enough to be interacting with it, and those who are more distant but who might be interacting with the potential partners.

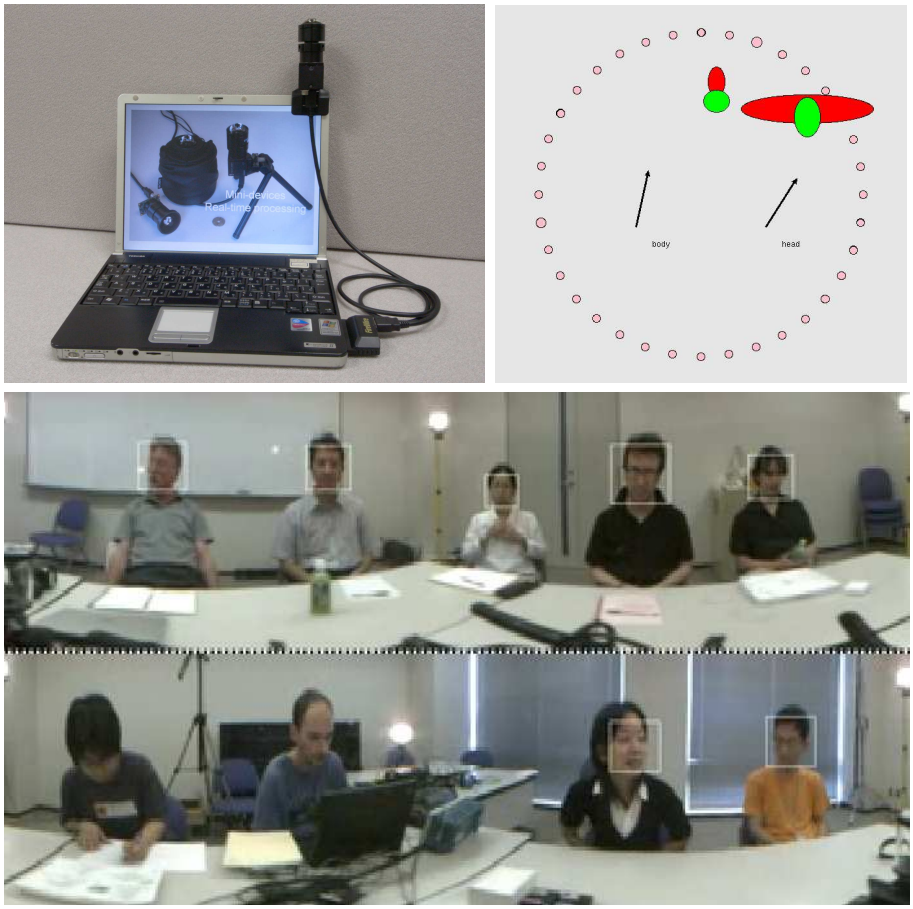
When someone close is detected, the system starts tracking the movement and activity associated with that person or persons. At the same time it also tracks differences in the audio environment and maps the two to derive added information such as (a) who is speaking, (b) the timing of their facial activity with respect to events in the speech, and (c) linking of nonverbal behaviour to non-lexical sounds in the speech.

The device is compact, unobtrusive, and inexpensive, yet is able to process data robustly in real-time, at a frame rate of up to 5 frames per second approximating the rate of syllables in the speech.

2 Summary of the technical content of the demonstration

The system links input from the video camera with that from an audio microphone. The software continuously measures the audio background level and notes any changes that might be from speech-related movements of a person or persons interacting with the system. It then attempts to distinguish between continuous speech, which is sent to the recogniser along with related activity information, and short bursts of non-lexical speech activity which are processed separately for their nonverbal content.

The output of the sensor system is a stream depicting the number of persons present within interacting range, their movements and activity, and an annotation of their verbal/nonverbal speech activity. For short non-verbal bursts of speech, the system uses spectral and prosodic information to estimate affective states, and for longer verbal speech, provides a time-aligned guide to facial activity during the speech.



3 Short storyboard describing the system

- The top left image shows the complete system consisting of a camera mounted on a small laptop computer. A key point of this demo is that sophisticated real-time processing is taking place in a robust manner on such hardware.
- The lower image shows a scene where nine participants at a business meeting are being tracked. The information inferred from the pictures is combined with the audio signal to provide a composite analysis of the scene. The system is able to reliably detect people participating in the scene and to ignore those who are not active at any given moment.
- The top right image shows such an analysis, where two people are determined to be active, and the person on the right more so than the one on the left. The row of dots mark peripheral activity around the edges of the area. The meters graphically indicate degree of activity. Intensity and direction of head motion is shown using red ovals, and body movement in green.

4 Background to the system

As described in [1], 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 is the 'Edge Sense II' from [2]. This algorithm provides good quality output while still being able to run at a reasonable speed. Other algorithms have been used [3] [4] but did not provide a significant advantage while being considerably slower. At this point the image still consists of a circular band and must be rectified before the face detection. To this end a simple linear resampling is performed. The resulting rectified image is now ready to be used for the face detection.

We use the Viola-Jones face detection method [6] [7] which is based on pattern matching, trained on a large number of images in the form of Haar cascades. 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.

The tracking can drift in time so it is limited by use of three safeguards. The first consists of limiting the time during which the gap-bridging tracking will be performed. 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 block-matching algorithm. The combination of the Viola-Jones detection, the face tracking and the above mentioned checks leads to 97% of faces detected during one hour meetings recorded both indoor and outdoor.

In parallel to the video processing, the audio level in the environment of the camera is measured using the internal microphone of the computer. Local

differences from the background level are then compared with facial and bodily activity measured from the detected participants to produce an estimate of who is talking, and then from features of the audio timing, whether the content is verbal or non-verbal.

Long stretches of speech activity are taken to be indicative of propositional content or deliberate information transfer, and therefore not of particular interest with respect to nonverbal processes, but the shorter speech utterances are processed further according to acoustic characteristics and timing coincidences.

The audio processing incorporates a non-verbal utterance detector based on SVM categorisation of the principal components of fourteen prosodic and spectral features of the shorter utterances to make use of intonation and voice-quality in categorising different affective uses of common interjections and backchannel utterances.

For simultaneous group reactions such as laughter, grunting, or nodding, further use is made of the timing relations and coincidences between sound bursts and bodily movements for the detection of topic boundary candidates.

The system is still in early stages of development, but a prototype version will be demonstrated for evaluation at the conference.

5 Acknowledgements

This work was funded under the SCOPE initiative #041307003. The first author is supported by NiCT, the National Institute for Communications and Information Technology. Both are under the Japanese Ministry of Internal Affairs and Communications,

References

1. N. Campbell and D. Douxchamps, “Processing Image and Audio Information for Recognising Discourse Participation Status through Features of Face and Voice”, in Proc Interspeech 2007.
2. T. Chen, “A Study of Spatial Color Interpolation Algorithms for Single-Detector Digital Cameras”, http://www-ise.stanford.edu/ tingchen/.
3. 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.
4. 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.
5. 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.
6. 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.
7. P. Viola and M. Jones, “Robust Real-Time Face Detection”, International Journal of Computer Vision, vol. 57, no. 2, pp. 137-154, May 2004.