Robust Real Time Face Tracking for the Analysis of Human Behaviour

Damien Douxchamps[1] and Nick Campbell[2]

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

Abstract. We present a real-time system for face detection, tracking and characterisation from omni-directional video. Viola-Jones is used as a basis for face detection, then various filters are applied to eliminate false positives. Gaps between two detection of a face by the Viola-Jones algorithms are filled using a colour-based tracking. This system reliably detects more than 97% of the faces across several one-hour videos of unconstrained meetings, both indoor and outdoor, while keeping a very low false-positive rate (<0.05%) and without changes in parameters. Diverse measurements such as head motion and body activity are extracted to provide input to further research on human behaviour and for tracking participant activites at round-table meetings and similar discourse environments.

1 Introduction

The analysis of the relation between human behaviour and speech has been the subject of numerous research in the past and has recently formed the core of integrated research on meetings activity. One particular case of interest is the analysis of discourse processes and human interactions in meetings because those are common, easy to setup and provide a relatively controlled environment while encouraging people to express themselves [1,2,3,4]. However, the various approaches used to track the people's behaviour in these circumstances often use intrusive equipment, like individual cameras and microphones. As intrusions into the discourse will inevitably change the behaviour of people, less invasive techniques are sought [5].

In this context, we have developed a real-time video system that relies on a single small omnidirectional camera to retrieve information about the attendants' motion and activity level. No specific lighting is required. Given the relatively low resolution of our video it is not possible to extract fine information such as eye gaze but we can still detect heads and calculate the person's motion and activity. In a later stage only briefly mentioned here, this data is then correlated with verbal and non-verbal speech to infer higher level information about behaviour of discourse participants.

A. Popescu-Belis, S. Renals, and H. Bourlard (Eds.): MLMI 2007, LNCS 4892, pp. 1–10, 2007.
© Springer-Verlag Berlin Heidelberg 2007

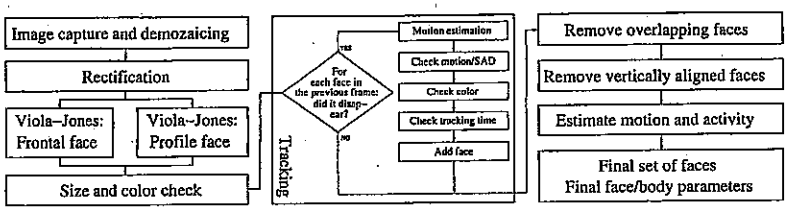

Fig. 1. Processing flow for each frame of the video

The detection and tracking of faces is a well covered subject in the literature. Among the different techniques available, the one proposed by Viola and Jones [6] [7] has the best results for low-resolution video and has been used in many situations. However, the number of studies reporting on their actual detection rate is surprisingly small, or they are limited to short video sequences. Examples include Fröba (90% detection rate, 0.5% false positive rate) [8], Kawato (89%, 1%) [9] and Castrillón-Santana [10]. In this paper we will show how to achieve a very high detection rate (>97%, <0.05%) in the case of unconstrained meetings lasting over an hour.

2 Video Processing

The processing techniques (Fig. 1) used in our system are standard and well documented, such as face detection and block matching (BMA). However, it is not trivial to build a real-time processing chain from these building blocks, especially when a high level of detection is to be achieved without any constraints given to the participants of the meeting.

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 as follows:

2.1 Video Capture, Demozaicing and Rectification

The video signal from a digital camera is decoded from a raw Bayer format to a full RGBI image. The demozaicing is performed using the 'Edge Sense II' algorithm presented in [12]. This algorithm provides good quality output while still being able to run at a reasonable speed. Other algorithms have been used [13] [14] but did not provide a significant advantage while being considerably slower. The circular, 360° image of Fig. 2 is then rectified with a linear subpixel

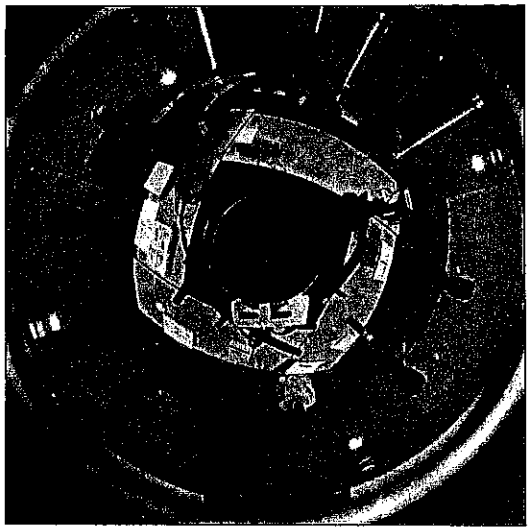

Fig. 2. Circular 360 degrees image captured by the camera (1040×1040)

Fig. 3. Rectified 360 degrees image (2048×260). Note the limited vertical resolution.

resampling before the face detection. To ensure a proper scaling of the faces, the horizontal size of the rectified image is set around $2\pi r$ where r is the average radial position of faces in the circular image. This rectification is necessary because the Viola-Jones face detection cannot detect faces in any orientation without a significant additional computational cost or a loss in accuracy. The resulting rectified image (Fig. 3) is now ready to be used for the face detection.

2.2 Face Detection

Face detection can be performed in a number of ways. The first technique that we tried was based on background subtraction and colour segmentation [15]. It has the advantage of not requiring an image rectification, but it failed to provide satisfactory results due to illumination changes and colour variability. A better approach is to use the Viola-Jones face detection [6] [7] which is based on pattern

matching. One drawback of this approach is that the algorithm must be trained on a large number of images, but standard software packages such as OpenCV [11] 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 50% of faces can be found during our round-table meetings.

To filter out the few non-faces that were detected we use two filters. The first one limits the size of the head within a reasonable range. The second one verifies that the face region contains a minimum of 25% of skin-coloured pixels. We found that the skin tone could be defined with sufficient accuracy in the RGB color space using the following criteria: $0.55<R<0.85$, $1.15<R/G<1.9$, $1.15<R/B<1.5$ and $0.2<(R+G+B)/3<0.6$. We have successfully used the same criteria for both indoor and outdoor scenes. At last, the binary mask of skin-coloured pixels is eroded and dilated using mathematical morphology before counting the number of skin-coloured pixels.

After this filtering, overlapping face-regions can still exist but they are removed easily by verifying that their overlapping region is not greater than around 20%. If so, the smallest face is discarded. Note that removing all faces that have the slightest overlap is not appropriate because people may be approaching each other for talking discretely, and their face regions may thus intersect slightly.

The Viola-Jones face detection is strictly frame-based. The lack of time integration means that the detection is not guaranteed to be continuous. In fact it can oscillate even with small image variations: a face can be detected in one frame, disappear in the next frame and then reappear again. To avoid these instabilities, we introduce a method to track the faces and bridge the gaps between two successive detections.

2.3 Face Tracking

If a face region in one frame intersects with a face region in the next frame, they will be considered to be from the same physical face and tracking is not necessary. If no such face can be found in the next frame, we will attempt to bridge the Viola-Jones detection gap by looking for an instance of the older face in the newer frame, using a classic block-matching algorithm (BMA) based on the Sum of Absolute Differences (SAD). This matching 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 will be performed. Given the very low false-positive rate of the face detection (see Section 3) we can allow a long maximal tracking time of 30 seconds. Secondly, we verify that the tracked face still contains a minimal amount of skin-coloured area, as we did after the Viola-Jones detection. Thirdly, the image difference between the old and tracked faces should be limited b y a threshold. 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

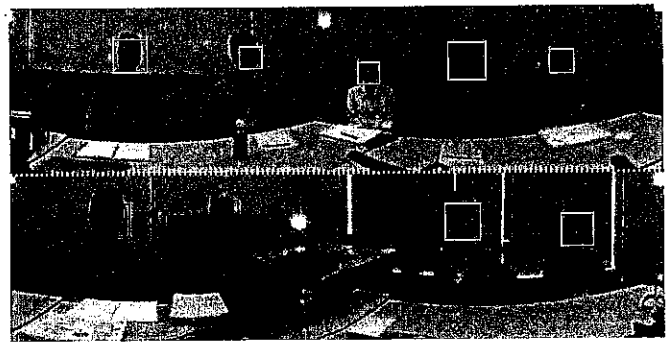

Fig. 4. Typical output from the program showing two 180 degrees sections on top of each other. Detected faces are shown with a white or black square.

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. A visual output of a final set of detected faces is shown in Fig. 4.

2.4 Motion and Activity Estimation

Once faces are properly detected a number of measurements are performed to identify their position, motion, and surface. 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. Two measures of a person's activity are also computed as the mean SAD between the previous and the current image: one is computed on the face region, and the other on the body region, the latter being defined as the area below the face with a width three times that of the face.

The graphs in Fig. 5 show a small section of five minutes of a few head and body measures for the nine persons attending a 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 in the vertical or horizontal head motion. Activity measures also correlate well with the global movements of a person. These measurements are now being correlated more systematically with manually labeled audiovisual data to provide clues about the link between physical activity and both verbal and nonverbal discourse events. This work, however, is beyond the scope of the present paper.

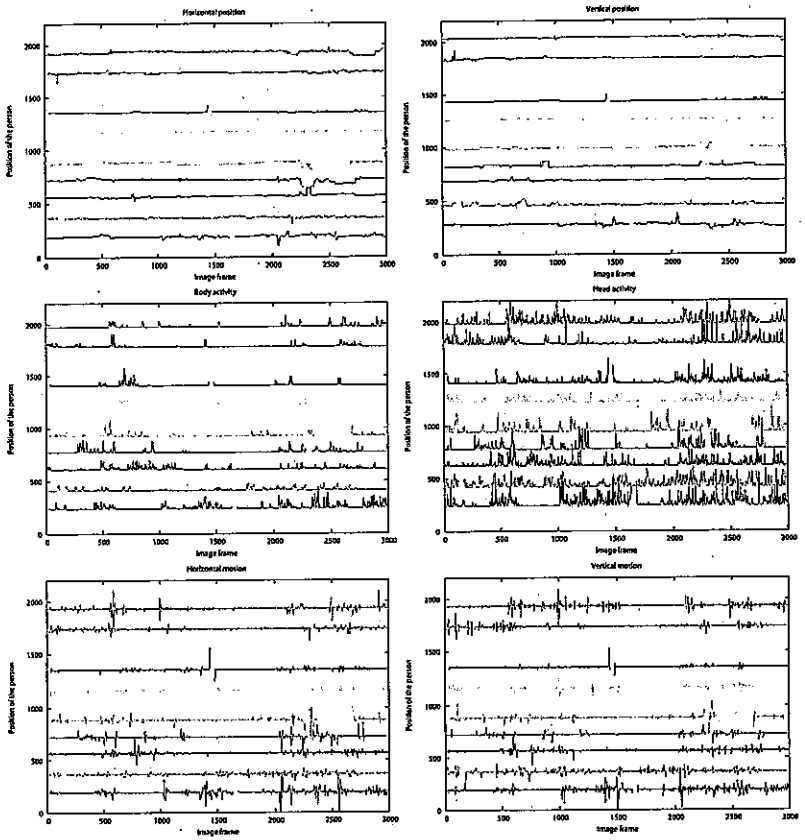

Fig. 5. The body and head activity (*top row*) and the head horizontal and vertical movement (*bottom row*) of the 9 participants found in the third sequence

3 Experiments

The proposed system has been tested under several conditions of lighting, image resolution and human activity. The minimum resolution for the circular image was found to be around 1000×1000; of the tests meeting this criteria five will be shown and discussed here. All tests were performed with the same hardware and processed with the same software and parameters. Typical output images are presented for each sequence in Fig. 6 together with a histogram of the frames of each sequence according to the number of faces detected in each of them. Ideally, each histogram should show a single bar for the bin corresponding to the number

of people attending the meeting. Due to various errors, detection will sometimes fail and frames with fewer people detected than expected will exist. Similarly, false positives may lead to frames with a higher number of detected people than expected.

The first sequence is a one-hour meeting recorded at 10 frames per second. It has a raw detection rate of 95%. Some of the faces are not present at all during some periods of time, for instance when a person leaves her seat to write on a white-board. If we take these long events into account the detection rate climbs to 97%. The unrestricted movements of people also leads to other numerous small undetectability events that are more difficult to take into account, such as looking at the ceiling, looking back, face obscured by a sheet of paper, and so on .... The 97% figure may thus be an underestimate. At the same time, the amount of false positives is less than 2%.

The second sequence shows limits in our approach, with a poor detection rate of 57%. This is explained by three factors: 1) a high contrast video with strong shadows is not optimal for our Viola-Jones detection; 2) the rectangular table means that people far from the camera will appear too small, which is also difficult for the Viola-Jones algorithm to detect and 3) the sharpness of the sequence was poor, washing face features away. Consequently, further tests were performed with a lower contrast and a square table, the latter leading to more homogenous face sizes than the rectangular table used in this test.

The third sequence has a similar detection rate to the first one: 96%. This test suffers a high false positive rate of 2% due to a high colour noise and poor white balance, as lights were switched on and off during the meeting to allow a video projection to be displayed. Many of the non-detections during this sequence are due to people looking away from the camera at table centre towards the presentation screen instead. Their faces are then strongly tilted or hidden, presenting angles that the Viola-Jones algorithm was not trained to detect.

The fourth sequence was taken outdoors while using identical processing parameters. Surprisingly, the detectability is also good (93%) but could without doubt benefit from fine tuning of the parameters. However, the strong directionality of the sunlight results in a set of brighter faces (facing towards the sun) and darker faces (back to the sun) which cannot be simultaneously optimized. This difference in exposure poses problems both for the Viola-Jones and for the colour-tracking.

Finally, the fifth sequence has a poor detection rate of 33%. The conditions were that of the second test but the camera was in the way of the projector beam which strongly influenced the white balance, lowered contrast and added a significant amount of flare. The histogram shows two distributions which correspond to the projector being on or off. The large number of tilted heads also partially explain the lower detection rate.

Overall, the face detection appears to work very well even in very different situations, provided that the scene has a reasonable focus, white-balance, sharpness and dynamic range.

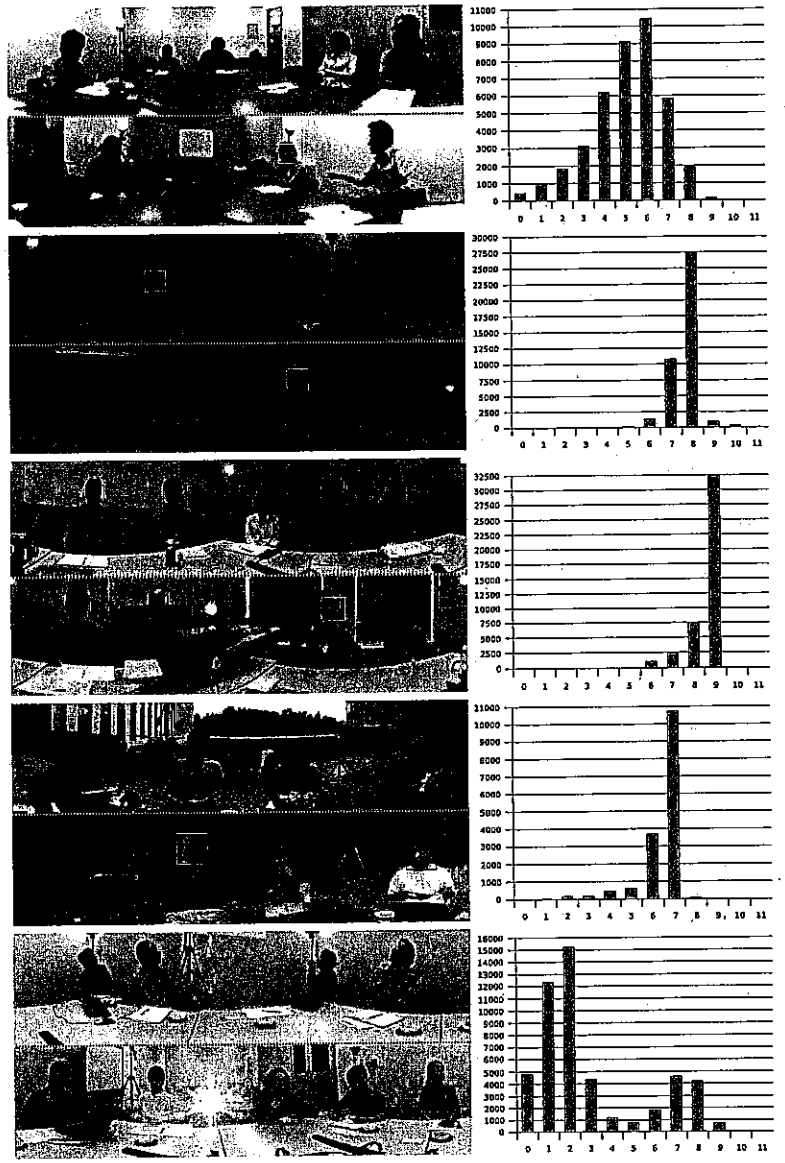

Fig. 6. Detection results for five test sequences. A representative image from the sequence is shown in the (*left column*). The (*right column*) contains the histogram of the number of images (*vertical axis*) per number of people detected (*horizontal axis*).

4 Conclusions

We have presented an image processing technique that is able to reliably extract faces from hour-long recordings of unconstrained meetings. Our technique is able to achieve a very good detection rate (>95%) while keeping the false positives to a negligible level (<1.5%). Conditions for which our approach has problems have been identified but can be easily avoided so that they do not limit its scope of application.

Acknowledgements

The second 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,

References

1. Dielman, A., Renals, S.: Multistream Recognition of dilogue acts in meetings in Renals. In: Renals, S., Bengio, S., Fiscus, J.G. (eds.) MLMI 2006. LNCS, vol. 4299, pp. 178–189. Springer, Heidelberg (2006)
2. Burger, S., MacLaren, V., Yu, H.: Meeting corpus: The impact of meeting type on speech style. In: Proc. International Conference on Spoken Language Processing (ICSLP), Denver (September 2002)
3. Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, C.: The ICSI meeting corpus. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Hong-Kong (April 2003)
4. McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., Zhang, D.: Automatic analysis of multimodal group actions in meetings. IEEE Trans. on Pattern Analysis and Machine Intelligence 27(3), 305–317 (2005)
5. Campbell, W.N.: A multimedia database of meetings and informal interactions for tracking participant involvement and discourse flow. In: Proc. LREC 2006, Genoa, Italy (May 2006)
6. Viola, P., Jones, M.: 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 (2001)
7. Viola, P., Jones, M.: Robust Real-Time Face Detection. International Journal of Computer Vision 57(2), 137–154 (2004)
8. Froöba, B., Küblbeck, C.: Face Tracking by Means of Continuous Detection. In: Proc. of the IEEE 2004 Conf. on Computer Vision and Pat. Rec. Workshops (CVPRW 2004), 27, 02, 65–65 (June 2004)
9. Kawato, S., Tetsutani, N.: Scale Adaptive Face Detection and Tracking in Real Time with SSR filter and Support Vector Machine. IEICE - Transactions on Information and Systems, E88-D 12, 2857–2863 (2005)
10. Castrillón-Santana, M., Déniz-Suárez, O., Guerra-Artal, C., Isern-González, J.: Cue Combination for Robust Real-Time Multiple Face Detection at Different Resolutions. In: Moreno Díaz, R., Pichler, F., Quesada Arencibia, A. (eds.) EUROCAST 2005. LNCS, vol. 3643, pp. 398–403. Springer, Heidelberg (2005)

11. OpenCV, http://www.sourceforge.net/projects/opencvlibrary
12. Chen, T.: A Study of Spatial Color Interpolation Algorithms for Single-Detector Digital Cameras, http://www-ise.stanford.edu/tingchen/
13. Hirakawa, K., Parks, T.W.: Adaptive Homogeneity-Directed Demosaicing Algorithm. IEEE Trans. on Image Processing 14(3), 360–369 (2005)
14. Chang, E., Cheung, S., Pan, D.: Color filter array recovery using a threshold-based variable number of gradients. In: Proc. of the SPIE Conference, vol. 3650, pp. 36–43 (1999)
15. Hsu, R.L., Abdel-Mottaleb, M.: Face Detection in Color Images. IEEE Trans. on Pattern Analysis and Machine Intelligence 24(5), 696–706 (2002)