MULTRAK: A SYSTEM FOR AUTOMATIC MULTIPERSON LOCALIZATION AND TRACKING IN REAL-TIME

O. Bernier, M. Collobert, R. Feraud, V. Lemaire, J. E. Viallet and D. Collobert

France Telecom CNET LAN/DTL/DLI
2, av. Pierre Marzin 22307 LANNION Cedex FRANCE
e-mail:olivier.bernier@cnet.francetelecom.fr

ABSTRACT

A real time system is described for automatic detection and tracking of multiple persons, in the context of video-conferencing systems. This system, called MULTRAK (MUltiperson Locating and TRacking Automatic Kernel), is able to continuously detect and track the position of faces in its field of view. The heart of the system is a modular neural network based face detector, giving fast and accurate face detection.

1. INTRODUCTION

One drawback of the current video-conferencing systems is their passivity. It limits the freedom of movement of the participants, and forces their intervention in case of a changing context: for example, the field of view of the camera may need a manual adjustment when a new participant arrives. The participants of a video-conference are at the same time the actors of the communication, but also, in a limited way, the cameramen and sound engineers.

To free the participants, we propose to create active and adaptive intelligent video-conferencing systems. The first step, with this goal in mind, is to build a subsystem capable of perceiving, in real time, the different positions of the participants. This feature can be used to adjust automatically the field of view of the broadcast camera, to control acoustic arrays for directive sound pick-up (see for example [2]), or to control a specialized camera for more focused views (of the current speaker for example).

Different systems locate and track faces in real time, using different techniques (see for example [4], [5], [6]) but in most cases, only one face is taken into account, and multiple detecting and tracking is not considered. Other systems tracking multiple moving objects exist, but do not check that what is tracked corresponds to a person or face (see for example [3]).

We present MULTRAK (MUltiperson Locating and TRacking Automatic Kernel) which locates and tracks the different persons of a video-conference. This system is based on the LISTEN system [2].

2. THE LISTEN SYSTEM

MULTRAK is derived from the LISTEN system, using the extended face detection subsystem described in [1]. Using skin color and movement detection, LISTEN is able to extract regions of interest. A neural network based face detector is applied on these regions, and once a face is detected, the system tracks the corresponding skin color region, controlling the pan, tilt and zoom of a motorized camera.

The use of a neural network to detect faces was proposed by several authors, and gave the best results published so far for face detection (see [7]). A previously trained neural network classifies a fixed size sub-image as face or non-face. Without prior knowledge, all possible positions and scales must be tested. The drawbacks of such an approach are its computational cost, and the need of a very low false alarm rate.

The size of the sub-image, tested by our neural network [1], is 15 by 20 pixels corresponding to 300 inputs. The modular architecture enables the detection of turned faces (up to 50 degrees) with good detection rate, and with a very low false alarm rate ($\approx 10^{-6}$). It is composed of two modules, one for front viewing faces, and one for turned faces. A multilayer perceptron gate network is trained to combine the results of the two previous networks. Our modular network obtained the best results published so far on the CMU test set A.

The use of skin color and movement detection reduces drastically the number of positions and scales to test: detection of a face in a region takes around 0.5s. Once a face is detected, the corresponding skin color region is tracked at 25 images per second. As our neural network has a very low false alarm rate, the tracked region is almost guaranteed to correspond to a face (for a region, the false alarm rate is around $5\,10^{-4}$).

3. MULTRAK

The MULTRAK system uses a fixed camera with wide field of view to analyze the scene. The image broadcasted is obtained by a second camera, with a frame controlled by MULTRAK. Using the same techniques used by LISTEN, the regions of interest (moving skin color regions) are extracted. Every new image, a different region of interest is tested with the face detector. For each detected face, the corresponding region is tracked. This region will no longer be tested for face detection, to save processing time.

All the regions corresponding to faces are tracked continuously. Heuristic criteria, based on the variation in appearance of the skin color region, are used by the system to estimate when the tracking of a region is faulty. In this case, this region is no longer tracked but will reappear as a region to be tested: a lost but still present face will quickly be detected again.

As only one region is tested for each new image, we obtain a constant processing rate, and as the tested region is different for each image, incoming new faces, corresponding to new regions of interest, are rapidly detected.

4. THE NEURAL NETWORK

To be efficient, the system needs a detection module faster than the modular neural network previously used with LISTEN, without degradation of the performances.

The idea is to create a much simpler neural network, called pre-network. It is a fast network with a very high detection rate (around 99%) but with a high false alarm rate (below 10 %). This network is a filter which discards more than 90 % of the hypothesis (location and scale of face) presented to it. The modular architecture previously reported is then applied to the remaining hypothesis. As the detection rate of the pre-network is very high, the detection rate of the combined system is close to the detection rate of the modular network alone. Since more than 90 % of the hypothesis are tested using only the pre-network, which is far simpler than the modular one, the combined system is faster.

The pre-network is a single Multi-Layer Perceptron (MLP) using the same input as the modular network described above. It has 20 hidden neurons, and one output (face/non-face), for a total of around 6000 weights (as compared to more than 70000 for the modular network). The pre-network is trained using standard back-propagation and the same face examples used for the modular network ($\approx$ 7000 front view and turned faces) and 24000 non-face examples. Some 7000 face examples and $\approx$ 24000 non-faces are used for the validation set to stop the training.

On the validation set, the pre-network achieves a detection rate of more than 99% for a false alarm rate below 3%. The combined system is tested on the CMU test set A ([7]). It gives an overall detection rate of 80% for a false alarm rate below 10^{-6}, as compared to 85% (with approximately the same false alarm rate) obtained by the modular network alone. The combined system detects turned faces up to 50 degrees (with detection rate over 75 %), and is four time faster. The lower detection rate is not critical, as with MULTRAK, if a face is not detected within a region of interest, this region will be tested again on future images.

5. DISCUSSION

The MULTRAK system is able to detect and track any number of persons, provided these persons are not too far from the camera (see figure 1). The size of the sub-image processed by the neural network corresponds to the minimum detectable face size. Combined with the size of the processed image (CIF format, 384 by 288 pixels), it gives a maximum field of view corresponding to three or four persons behind a table in the context of a video-conference. The extension of this field of view is possible provided a complete image (PAL format, 768 by 576 pixels) is processed, giving a field of view suitable for up to seven persons. The system operates at a constant rate of around 7 images per second. The detection of a new participant is typically achieved in less than one second.

The current limitations of the system are the problem of backgrounds with skin color, which confuses the tracking algorithm, and the problem of region crossings. Two crossing regions are merged into one, and in this case, the regions appearance is not stable. The two regions are not tracked anymore, until a face is detected (inside the merge region, or when the two regions are separated again).

6. REFERENCES

[1] R. Feraud, O. Bernier, J.E. Viallet, M. Collobert and D. Collobert, "A Conditional Mixture of Neural Networks for Face Detection, Applied to Locating and Tracking and Individual Speaker", in *Proceedings of the 7th International Conference on Computer Analysis of Images and Patterns, CAIP'97*, Kiel, Germany, 1997.

[2] M. Collobert, R. Feraud, G. Le Tourneur, O. Bernier, J. E. Viallet, Y. Mahieux, and D. Collobert, "LISTEN: A System for Locating and Tracking Individual Speakers", in *Proceedings of*

Figure 1: Left: image analyzed by MULTRAK. Right: Corresponding automatically framed image. Top: Videoconference with two persons. Bottom: arrival of a new participant. White rectangles indicate the faces tracked.

the second International Conference on Automatic Face and Gesture Recognition Killington, Vermont, October 1996.

[3] S. Intille, J. Davis and A. Bobick, "Real-Time Closed-World Tracking", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR'97*, Puerto Rico, June 1997.

[4] M. Hunke, and A. Waibel, "Face Locating and Tracking for Human-Computer Interaction", in *Proceedings of the 28th Asimolar Conf. on Signals, Systems, and Computers*, Pacific Grove, California, November 1994.

[5] N. Oliver and A. Pentland, "LAFTER: Lips and FAce Real Time Tracker", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'97*, Puerto Rico, June 1997.

[6] A. Eleftheriadis and A. Jacquin, "Automatic face location detection and tracking for model-assisted coding of video teleconferencing sequences at low bit-rates", in *Signal Processing, Image Communication*, Vol 7, 1995.

[7] H. Rowley, S. Baluja, and T. Kanade, "Human Faces Detection in Visual Scenes", in *Advances in Neural Information Processing Systems 8*, 1995.