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

LAN/DTL/DLI
France Télécom CNET
2, av. Pierre Marzin 22307 LANNION Cedex FRANCE

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 current video-conferencing systems is their passivity. It limits the freedom of movement of the participants, and requires 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. With this goal in mind, the first step is to build a subsystem capable of perceiving, in real time, the positions of the different participants. This feature can be used to adjust automatically the field of view of the camera used to obtain the image broadcasted to the distant participants. It can also be used 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. For example, using skin color detection possibly completed by shape analysis ([4]and [5]), using ellipses detection ([7]), or three dimensional analysis with two cameras ([6]). In most cases, however, 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), a system which locates and tracks the different persons participating in a video-conference, and is able to control a second camera for optimal framing. 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 in images was proposed by several authors, and gave the best results published so far for face detection ([8] and [1]). A previously trained neural network classifies a fixed size sub-window as face or non-face. Without prior knowledge, all possible positions and scales of a face in an image must be tested. The drawbacks of such an approach are its computational cost, and the need of a very low false alarm rate.

In LISTEN, face detection is based on the classification of a 15 by 20 pixels sub-window, after normalization, using a modular neural network. Normalization of the sub-window is done by histogram enhancing, smoothing and subtracting the average face. The modular network is composed of three networks. Two of these networks, one for front viewing faces, the other one for turned faces, are MultiLayer Perceptrons (MLP) with two hidden layers. They are trained to reconstruct the faces presented as their input. A non face sub-window is reconstructed as the nearest face in the estimated set of face learned by the network. The distance between a sub-window and this set of faces is then estimated by the distance between this sub-window and its reconstruction by the network. In this way, each network is able to estimate the probability that a sub-window is a face: a front viewing face for the first one, a turned face for the second one. The third gating network, a simple MLP with one hidden layer, is used to combine the outputs of the two networks. This modular architecture was tested using the standard CMU test set A [8]. The combination of the three networks give a good detection rate (85 %) with

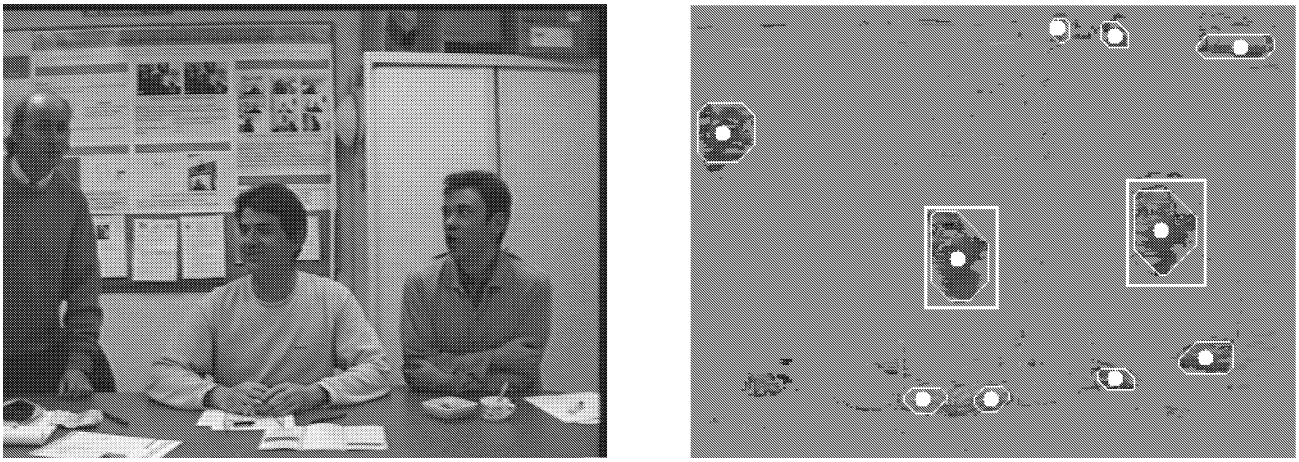

Figure 1: Left: image analyzed by MULTRAK, from the first camera. Right: the obtained regions of interest and two detected faces (white rectangles). The left person has just been entering the field of view and is not yet detected.

false alarm rate below 10^{-6}, which is the best result published so far on this test set. It is able to detect turned faces (up to 50 degrees) with a detection rate above 70 %.

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 about 0.5 sec. 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 approximately 5.10^{-4}, as 500 positions and scales are tested for each one).

3. The MULTRAK system

As opposed to LISTEN, which uses only one motorized camera for detection and framing of a person, MULTRAK uses a fixed camera with wide field of view to analyze the scene. The image broadcasted to the distant participants of the video-conference is obtained by a second camera, with a frame controlled by MULTRAK. Using the same skin color detection technique used by LISTEN, the regions of interest (skin color regions) are extracted. For every new incoming 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. Figure 1 shows an example of the detected regions of interest, as well as two faces tracked by the system.

All the regions corresponding to faces are tracked continuously. The regions of interest not corresponding to faces are tested again regularly, one for each new image acquired through the fixed camera. 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 reappears as a region to be tested. As only one region is tested for each new image, we obtain a quasi-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. A face lost by the tracking system but still present in the image is also quickly detected again.

One important difference between LISTEN and MULTRAK, is the constant application of the face detector to regions of interest not corresponding to a tracked face to detect possible new faces. As a consequence the face detector must be fast enough to be used for each image acquired. This is not the case of the face detector used by LISTEN.

4. The neural network

To be efficient, the MULTRAK system needs a detection module faster than the modular neural network previously used with LISTEN, without degradation of the performances. To obtain such a face detector, the idea is to create a much simpler neural network, called pre-network. It is a relatively small and fast network with a very high detection rate (above 95 %) but also with a high false alarm rate (up to 3 %). This network, unusable alone because of its poor false alarm rate, can be used as a filter which discards more than 97 % of the hypothesis (location and scale of possible face) presented to it. This filter is used after the normalization on the extracted

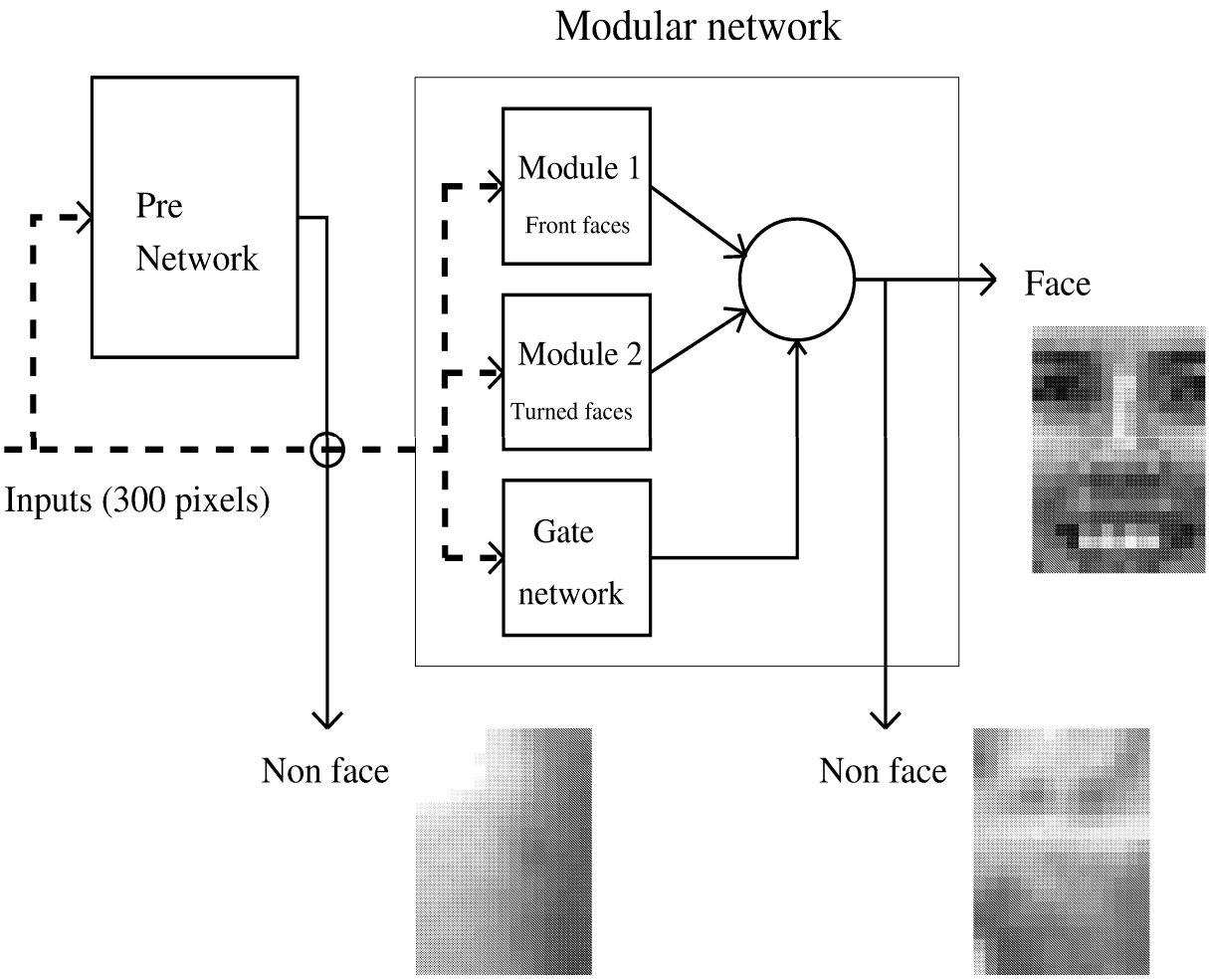

Figure 2: General structure of the combined face detector

sub-window corresponding to the hypothesis (15 by 20 pixels). The normalization is the same as the one used in LISTEN (histogram enhancing, smoothing and subtraction of the average face). The modular architecture previously reported is then applied to the sub-windows corresponding 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 97 % 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 (15 by 20 pixels). 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 ($\approx$ 14000 front view and turned faces). These examples represent faces centered in the 15 by 20 pixels sub-window and at a scale corresponding to its size. Approximately 50000 specific non-face examples (15 by 20 pixels sub-windows, which do not correspond to faces) were obtained using an iterative algorithm. A first pre-network is trained using a certain number of sub-windows collected randomly in natural images without faces as non face examples. This pre-network is then tested on natural images and a random subsample of the false alarms of this pre-network are added to the set of non-face examples for a new training. The obtained pre-network is tested again and random false alarms are added again to the set of non-face examples. This process is repeated until the false alarm rate is satisfactory. At each stage, half of the face set and of the current non-face set are used as training sets. The remaining halves are used for the validation set to stop the training.

The general structure of the combined network is shown in figure 2. 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 ([8]). 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 a detection rate over 70 %). 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. Due to its size, the pre-network alone is approximately ten times faster than the modular network. However, approximately 3 % of the sub-windows must still be processed by this modular network. Taking into account the normalization time, the combined face detector is more than four times faster than the modular face detector. Even on regions corresponding to faces, the pre-network discards a great number of possible positions and scales, as it was trained to detect correctly scaled and positioned faces.

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 3). 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 (480 by 360 pixels), it gives a maximum field of view corresponding to four or five persons sitting 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 person is typically achieved in less than one second.

The current limitations of the system are the problem of background with skin color, and the problem of region crossings. The presence of skin color in the background, near the position of a face, can create a region of interest including both the face and a part of the background. In this case, detection by the network is more difficult. The presence of skin color in the background creates also a number of added regions, which increases the possibility of region crossing. Two crossing regions are merged into one, and in this case, the regions appearance is not stable. The heuristic used by the tracking algorithm can consider the face to be lost. If the two regions corresponded to a face, one is automatically lost, as only one region exists now whereas two where present before. In all cases, the system is able to recover a correct functioning when the two regions part, and the lost face is detected again in the following images.

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 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. Azarbayejani, C. Wren, and A. Pentland, "Real-Time 3-D Tracking of the Human Body", in *Proceedings of IMAGE'COM 96*, Bordeaux, France, May 1996.

[7] 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.

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

Figure 3: Left: image obtained through the fixed camera analyzed by MULTRAK. Right: Corresponding automatically framed image, obtained through the second camera. Top: Video-conference with two persons. Bottom: arrival of a new person. White rectangles indicate the tracked faces.