A system for tracking and recognizing pedestrian faces using a network of loosely coupled cameras

L. Gagnon*[a], F. Laliberté[a], S. Foucher[a], A. Branzan Albu[b], D. Laurendeau[c],
[a]R&D Dept., CRIM, 550 Sherbrooke Street West, Suite 100, Montreal, QC, Canada, H3A 1B9
[b]Dept. of Elec. & Comp. Eng., Univ. of Victoria, Victoria, BC, Canada, V8W 3P6
[c]Dept. of Elec. & Comp. Eng., Laval Univ., Quebec, QC, Canada, G1K 7P4

ABSTRACT

A face recognition module has been developed for an intelligent multi-camera video surveillance system. The module can recognize a pedestrian face in terms of six basic emotions and the neutral state. Face and facial features detection (eyes, nasal root, nose and mouth) are first performed using cascades of boosted classifiers. These features are used to normalize the pose and dimension of the face image. Gabor filters are then sampled on a regular grid covering the face image to build a facial feature vector that feeds a nearest neighbor classifier with a cosine distance similarity measure for facial expression interpretation and face model construction. A graphical user interface allows the user to adjust the module parameters.

Keywords: Video surveillance, face recognition, pedestrian detection, multi-camera network, facial expression

1. INTRODUCTION

This paper presents the technical characteristics and usage of a Face Recognition Module (FRM) that has been developed for a video surveillance system, called MONNET (MONonitoring of extended premises using a NETwork of loosely coupled cameras)[1]. The MONNET project aimed at developing an intelligent computer vision-based monitoring system for non-intrusive detection and recognition of persons moving around in public premises.

Detecting and characterizing a person is a difficult problem in non-controlled monitoring applications. It is required, for instance, in order to detect and identify suspicious behavior or to guide a person towards a given destination. In this context, several system architectures have recently been proposed and implemented (for instance, Ref. 2-4). In Ref. 2, the system allows a human operator to monitor activities over a large area using multiple calibrated cameras with a geospatial site model. Tracking approach is based on image correlation mapping, followed by computation of the 3D location on the site model. Inter-sensor communication consists in a "handing off" mechanism between sensors situated along the object's trajectory. A decentralized architecture is proposed in Ref. 3 using multiple calibrated cameras to learn patterns of activities from motion observation. The basic assumption for learning is the preservation of the object identity throughout the tracking process. Another wide area surveillance system using client-server architecture is proposed in Ref. 4. It uses noncalibrated cameras with overlapping and/or non-overlapping Fields of Views (FOVs). The system is trained to learn the topology of the FOVs. The inter-camera correspondence is established based on linear velocity prediction and on a spatio-temporal constraint based on the FOVs topology. Unlike most of the current systems[5], MONNET has a decentralized architecture; thus no dependence on a central server that could fail during an operational mode. Furthermore, the intelligent nodes send and receive information between them. Finally, infrared (IR) camera (in addition to visible ones) is attached to each node to improve performances in low-light conditions.

Since face recognition is not invariant to facial expressions, the face recognition performance can be improved by comparing faces having the same facial expression, thus requiring a step of facial expression recognition. Most facial expression recognition and interpretation algorithms[6] rely on controlled image/video acquisition conditions (person sitting and facing a camera, sufficient and uniform illumination, high resolution image/video). None of those are present here, resulting in a highly non-controlled environment. In MONNET, we classify facial expressions into basic emotions

* langis.gagnon@crim.ca; phone 514-840-1235; fax 514-840-1244; crim.ca

rather than using the more detailed Facial Action Coding System[7], because global face changes are more relevant for this application.

The paper is organized as follows. Section 2 gives a description of the global MONNET architecture. Section 3 gives a description of the architecture and use of the FRM. More scientific details about the implemented algorithms, experimental database and test performances can be found elsewhere[8].

2. MONNET ARCHITECTURE

The MONNET system is composed of a network of cameras sparsely positioned over an extended premise and connected to computing "nodes". Each node detects and track persons in the FOV of its cameras, characterizes these persons in terms of their appearance, broadcasts relevant information to other nodes on the network, and builds a log file describing the activity that has occurred in the monitored area[1].

Each node consists of four modules: acquisition, segmentation, tracking and person identification. The acquisition module gets video data from an IR and a standard video camera sharing the same FOV. The data are calibrated in temperature and in geometry. The segmentation module extracts the body into Region of Interests (ROIs). If only visible data are available, a pixel-based statistical foreground-background subtraction algorithm based on Gaussian mixture models (GMMs) is used[9]. When IR and visible data are available, a data fusion algorithm is used[10]. The tracking module aims at establishing coherent spatio-temporal relations between ROIs using a 5-point human model (head, hands and feet)[11]. The person identification module builds and updates an appearance model for each tracked pedestrian using color information from three parts of the body (head, torso, and legs)[12] and face biometrics[8]. The later is provided by the FRM. The model is sent to all other nodes for comparison and matching.

The physical configuration of each node consists in a main computer and few satellite ones (Figure 1). The main computer controls the video acquisition process and the optional graphical user interface (GUI). The computing thread may be distributed across the main computer and additional satellite computers if the computational load is too high for the main computing unit. Satellite computers may be used for the distributed implementation of a single module or for implementing several modules.

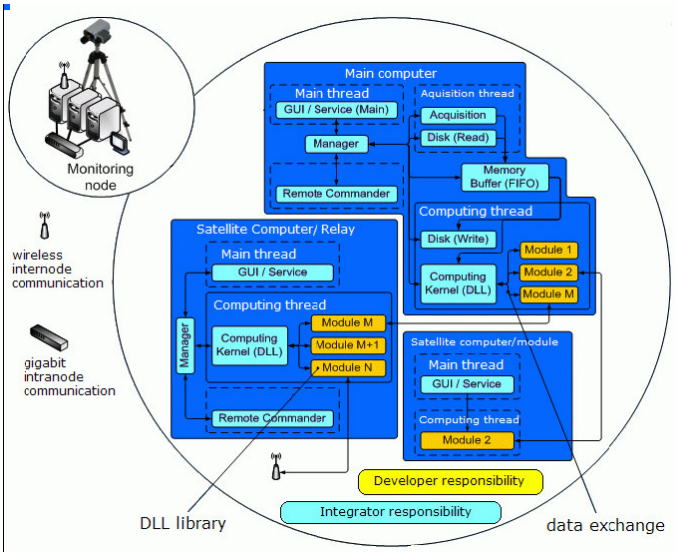

Figure 1: Physical configuration of a node in MONNET

An optional node can be added to the system for a graphical user interface to interact with the system and visualize the results. For example, it allows adding or removing modules that are not essential in order for the system to work. Such

modules include face recognition, clothes model, and body tracking. One can also change the underlying algorithm parameters. Results that can be visualized include the recognized facial expressions, recognized pedestrians, confidence levels, detected faces and facial features, body parts trajectories, background subtraction, visible and IR fusion, etc.

Finally, an inter-node communication kernel allows the nodes to coordinate their actions. While the network is running, nodes can be added/removed from the network and then start/stop receiving/sending appearance models to/from the other nodes by the use of a zero-configuration networking protocol (the Bonjour® protocol from Apple Computer Inc.). The MONNET system is thus configured as a wireless network of collaborative nodes, able to exchange information about the tracked pedestrians in an asynchronous manner.

3. FACE RECOGNITION MODULE

3.1 Architecture

The architecture of the FRM can be decomposed in six parts: (1) face detection, (2) facial feature detection (3) face normalization, (4) facial expression interpretation, (5) face model construction and (6) face model comparison (face recognition). The main elements are described in details in Ref. 8. We briefly summarize them in the following.

3.1.1 Face detection

Face detection is frame-based and done by a cascade of simple linear classifiers as implemented in the Intel® OpenCV library. Face dimension as small as 24 x 24 pixels can be detected, with a maximum profile view of $\pm 45^{o}$. A face frontal view selector, based on a spatial detection map of face candidates, is then applied in order to retain only good quality face images for the further facial feature detection step. To this aim, a frontal view quality factor is calculated which measures the fraction of maximal detections contained in a region centered on the center of mass of the spatial detection map of face candidates. A threshold is set on this frontal view quality factor to keep only frontal views.

3.1.2 Facial feature detection

Cascades of classifiers were also trained for the detection of nasal root, eyes, nose and mouth. Each training set was composed of about 3000 positive samples and 5000 negative ones taken from the BioID Face Database[13] which is manually annotated according to the FGNET standard[14]. (Additional public databases were also manually annotated by us for test purpose. Those are available online to the research community[15]). The nasal root detector is applied first. Once the nasal root is found, the face is separated in two regions. The eye detector is applied on the top region, the nose and the mouth detectors on the bottom one.

3.1.3 Face normalization

The face normalization part is based on the average anthropometric face proportions of a North American Caucasian human face (man or woman) of 19-25 years old[16]. The position and dimension of the four facial features detected previously are used to map the detected face X onto a normalized one X' using a two-step transform $X' = \mathbf{B}\cdot\mathbf{A}\cdot X$, where $\mathbf{A}$ is a similarity transform (from the detected face image to the anthropometric reference), and $\mathbf{B}$ is a combination of scale change and translation that maps the result to a normalized image of 129 x 209 pixels.

3.1.4 Facial expression interpretation

The facial expression interpretation follows five main steps: (1) Gabor filtering (with 40 different wavelets), (2) histogram equalization of the Gabor coefficients magnitude, (3) subsampling on a 26 x 42 grid, (4) PCA projection and (5) facial expression classification. At the end of step 3, each image is represented by a vector of dimension 43680. This is reduced to a vector $\mathbf{f}$ of dimension 50 after step 4. Step 5 uses a nearest neighbor assignment against a database of 523 expression-labeled images (about 75 images for each expression) coming from even public databases.

The choice of the classifier and distance measure has been done after an off-line performance test on a subset of 2739 images coming from public databases[8]. Two types of classifiers and two types of similarity measures have been tested: (1) nearest neighbor and template matching classifiers, (2) cosine and Euclidean distance measures. The nearest neighbor classifier with a cosine similarity measure has been chosen because it was performing the best for facial expression interpretation and face recognition[8].

3.1.5 Face models construction

The FRM is called each time a pedestrian is detected by the segmentation module. A body bounding box is sent to the FRM, as long as the body is in the camera FOV. The FRM checks if a frontal face is present in the upper part of the body bounding box on a frame-by-frame basis (no tracking). Each time the face is detected, a face feature vector **f** is extracted and used to initiate or update a face model of the person.

Two types of face models are calculated: dependent or independent of the facial expression. The expression-dependent face model is a set of at most seven vectors $\mathbf{u}_i = \sum_{j=1}^{N_i} \mathbf{f}_j$ (i = 1,…,7; for the 7 expressions), where N_i is an associated integers to $\mathbf{u}_i$ representing the total number of feature vectors accumulated for expression i since the beginning of the body detection. If N_i = 0, no vector exist for the corresponding expression i. At the first frame where a face is detected, the model is thus composed of only one vector (for the corresponding detected facial expression). Each time a new expression is detected for the same face, a new vector is initiated and added to the model. The face model is updated as long as the person is in the camera FOV. When the pedestrian leaves the FOV, the model is stored in the node database. A person can have more than one face model stored in one node database or in more than one node databases. The expression-independent model is constituted of only one vector and one associated integer, which is simply the addition of all feature vectors **f** extracted since the beginning of the body detection.

3.1.6 Face model comparison

Each time the current face model is updated, it is sent to the other nodes for comparison by the FRM based on a similarity measure c. It is also compared to the models stored in the same node.

Two measures have been implemented; one for each type of face model. For the expression-dependent type, we use the expression-weighted measure $c = \sum_{i=1}^{7} w_i * d(\bar{\mathbf{v}}_i, \bar{\mathbf{u}}_i) \Big/ \sum_{i=1}^{7} w_i$, where $\bar{\mathbf{u}}_i$ and $\bar{\mathbf{v}}_i$ are the two models, $d(\cdot,\cdot)$ is the cosine similarity measure and w_i are the weights. The weights are constants and given by the expression recognition rates calculated by cross-validation on the database of 523 images mentioned above (see Section 3.1.4). In this way, more confidence is put on the expressions that are easier to recognize than the others. For the expression-independent type, we use $c = d(\bar{\mathbf{v}}, \bar{\mathbf{u}})$. The similarity measure returns a number between 0 and 1 which is a confidence measure about the face match. This constitutes the output of the FRM which is sent to the MONNET controller for combination with other similarity measures obtained from the other appearance models.

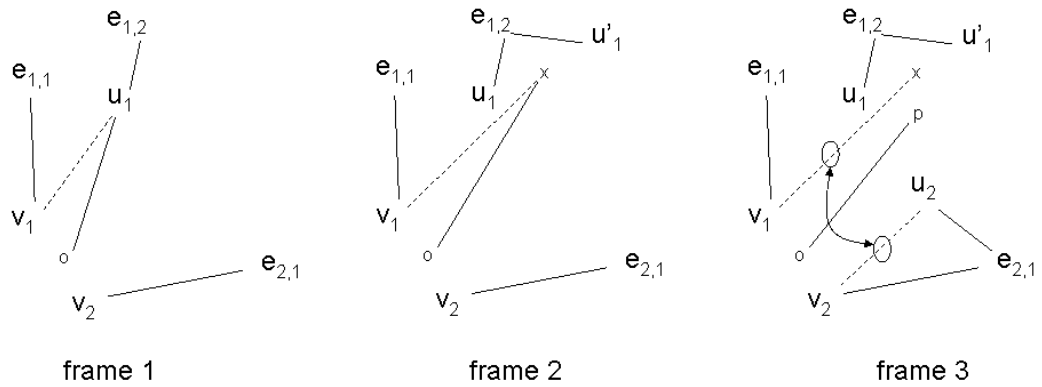

Figure 2: Schematic representation of the face model update and similarity measures (see text for details)

Figure 2 gives a schematic representation of the face model update to help understand the whole process. Suppose that initially, a node database has one face model stored in it, constituted of two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ for two different expressions. At frame 1, a face is detected and a feature vector $\mathbf{u}_1$ is calculated. This feature vector is classified as expression 1 (the point $\mathbf{e}_{1,2}$ represents the second representative of the expression 1 in the expression database, and is the

closest). The distance between the point $\mathbf{u}_1$ and $\mathbf{v}_1$ is the expression-dependent distance. The point **o** is the expression-independent face model in the database. The distance between the point $\mathbf{u}_1$ and **o** is the expression-independent distance. At frame 2, a feature vector $\mathbf{u'}_1$ is calculated which "moves" $\mathbf{u}_1$ to the point **x**. The distance between the point **x** and $\mathbf{v}_1$ is the expression-dependent distance. The distance between the point **x** and **o** is the expression-independent distance. At frame 3, a feature vector $\mathbf{u}_2$ is calculated which "moves" **x** to the point **p** but only for the expression-independent model. The weighted distance between the point **x** and $\mathbf{v}_1$ and $\mathbf{u}_2$ and $\mathbf{v}_2$ is the expression-dependent distance. The distance between the point **p** and **o** is the expression-independent distance.

3.2 Interface

Figure 3 gives a snapshot of the configuration window of the FRM. Three options are available in the **Detection** section. If **Features** is selected, the facial features (nasal root, eyes, nose, and mouth) are detected according to the selection made in the **Normalization type** section. No pedestrian comparison is made if this option is selected. If **Expressions** is selected, the facial features are detected, the facial expressions are recognized and the pedestrian is compared. **Expressions independent** is the same as **Expressions** but the face recognition does not take into account the facial expression. The **Interval time** parameter gives the interval before computing the facial expression and face recognition after one has been computed. This is to take into account the fact that the processing time is not real-time for now. The higher this value, the less face images is used for the face model.

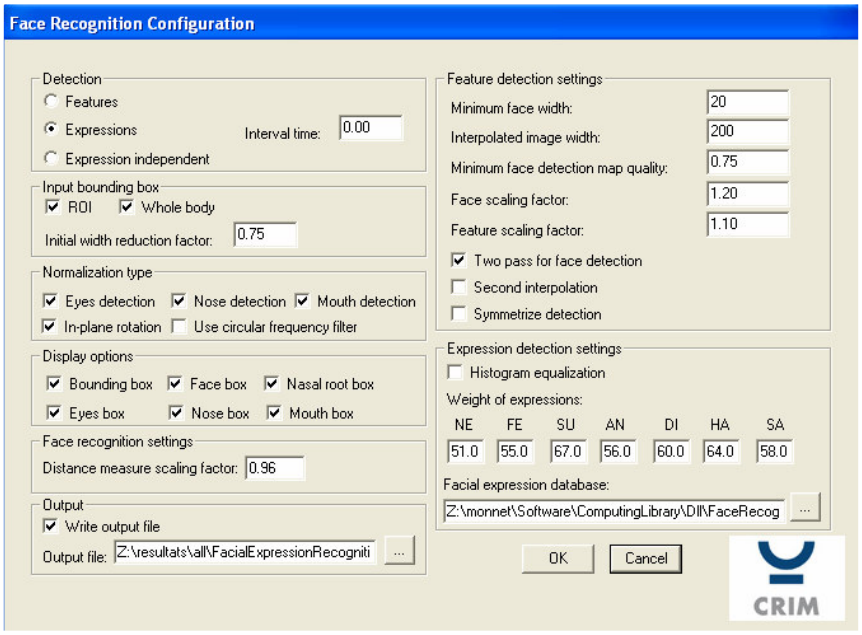

Figure 3: Configuration window of the FRM

The **ROI** option in the **Input bounding box** section is selected when a ROI is passed to the FRM, which is actually the case in the final MONNET system. The **Whole body** option is selected if the ROI sent to the FRM is the pedestrian body, which is also actually the case. The FRM resizes the ROI that it receives because for optimal face detection, the bounding box must be approximately centered on the head. The resized ROI is the same width as the original one but its height is equal to its width and it is translated up by 25% of its width. The **Initial width reduction factor** parameter represents the percentage of the width of the ROI that is actually used by the FRM. For optimal face detection, the width of the ROI for the head must be twice the head width. The parameter is used to obtain this optimal width. Left part of Figure 4 shows an example of a ROI that is passed to the FRM (large box) and the portion of this region that is actually used in the module (small box).

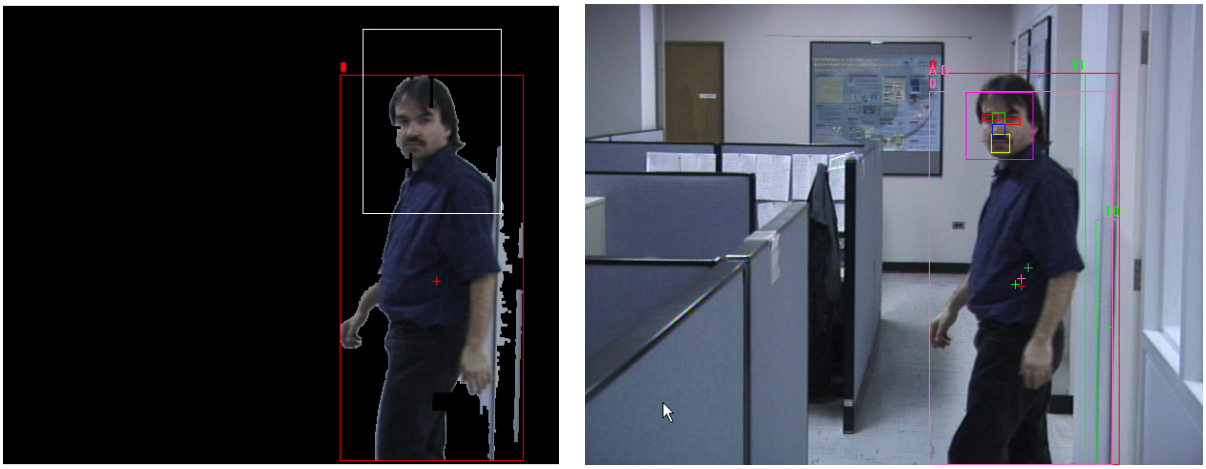

Figure 4: Example of the use of the **Whole body** option and the **Initial width reduction factor** (left) and display of the detected facial features as selected in the **Display** options section of the configuration window (right)

The five options in the **Normalization type** section sets the facial features used to compute the face normalization transformation. If **In-plane rotation** is selected, the normalization includes a correction for rotation in the plane. If **Use circular frequency filter** is selected, a circular frequency filter is used to find the orientation of the line connecting the eyes, which is used to compute the in-plane rotation. This option is enabled only if **In-plane rotation** is selected. The **Display options** section selects which detected features are going to be drawn over the visible image.

The **Minimum face width** parameter in the **Feature detection settings** section is used to reject the input bounding boxes and the detected faces that are too small. The **Interpolated image width** parameter is used to resize the bounding box prior to face detection (the classifiers have been trained for features of a given size). The larger this value, the longer is the computation time. The **Minimum face detection map quality** parameter is used to determine the minimum quality of the face detection (see Section 3.1.1). A smaller value will detect back of the head as valid faces. This value must be proportional to the image resolution. The **Face scaling factor** parameter sets the step in the face detection search region. The larger the step, the faster is the face detection but the less accurate it is. The **Feature scaling factor** parameter has the same use as the **Face scaling factor** but for the detection of the nasal root, the eyes, the nose, and the mouth. The **Two pass for face detection** option allows performing a first face detection on the image decimated by a factor of 2. It also allows performing more accurate face detection on the full resolution image if the first detected face is larger than the minimum face width and its quality is larger than the minimum face detection map quality. If **Second interpolation** is selected, a finer face bounding box (evaluated from the nasal root detection) is used for the detection of eyes, nose, and mouth. The **Symmetrize detection** option flips the image horizontally before performing a second detection (for the eyes, nose, and mouth) and takes the maximum of both detections.

The **Expression detection settings** section sets the options for the facial expression recognition and the pedestrian comparison. The **Histogram equalization** option allows to perform a histogram equalization of each Gabor filtered image. The **Weight of expressions** parameters contain each facial expression recognition rate computed with cross-validation on the facial expression database. The default values should not be changed unless the user wants to compare pedestrians using only one facial expression. In this particular situation, the weight of the desired expression should be set to 100 and all the other weights to 0. The default values must be changed if a different facial expression database is used. The **Facial expression database** field sets the file name containing the facial expression database that is used to recognize the facial expressions.

In the **Face recognition** settings section, the **Distance measure scaling factor** parameter is used to rescale the similarity between two pedestrians. This parameter must be increased if the value between two persons is not low enough. The **Output** section allows writing different results in a file for statistical analysis. Figure 5 gives an output example of the FRM showing detection, including face detection, facial feature detection, face normalization, and

expression interpretation. Figure 6 shows why the combination of different appearance models makes the system more robust. This is an example where the color appearance models fail but the face model succeeds.

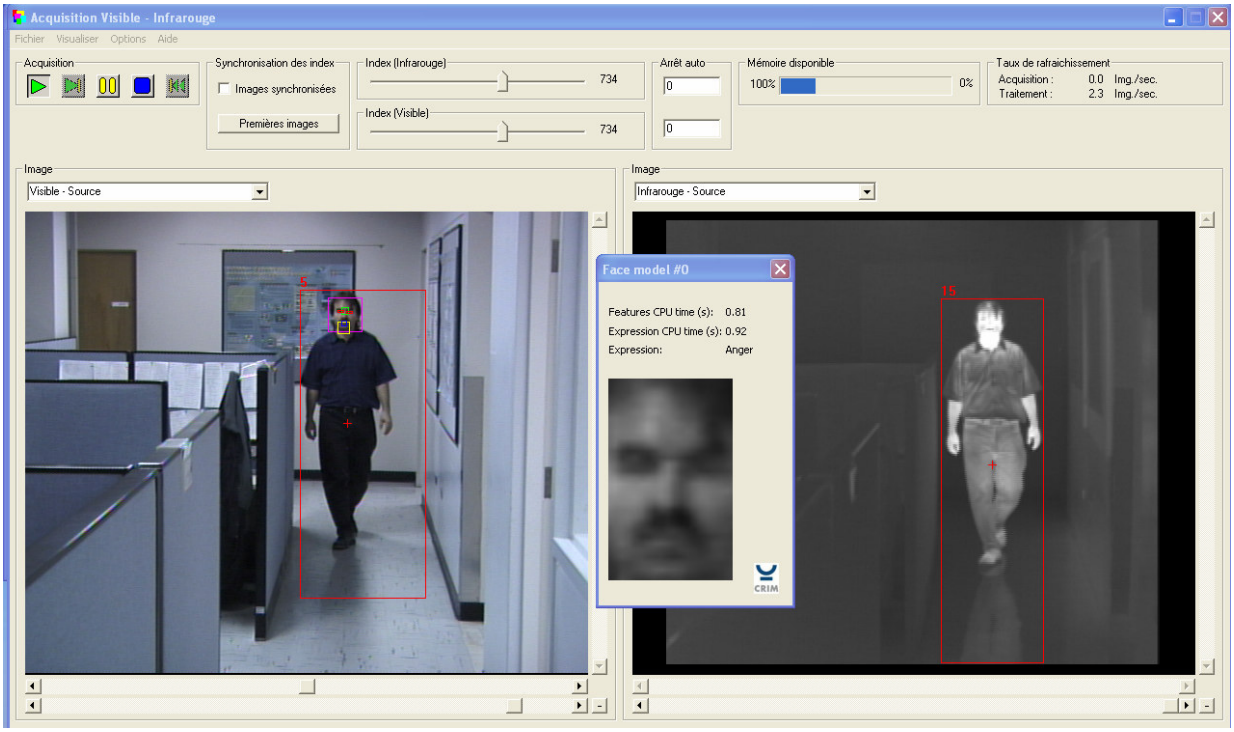

Figure 5: Output example of the FRM.

3.3 Performance

No extensive in-the-field tests have been done yet for the FRM. However, tests have been performed using public still image databases. First, face expression recognition performance has been measured on a dataset of 2739 images coming from seven public databases[8]. An overall facial expression interpretation rate of about 74% was obtained. The best recognized emotions were, in order, surprise, neutral, happiness, anger, sadness, disgust, and fear. Second, a face recognition performance of 75% was also obtained from a database of 1506 neutral face images of 812 subjects.

The MONNET system has been tested with up to five nodes (including the optional GUI one), wireless, in indoor environments (laboratory and conference demo room), during several days, and with a thousand pedestrians. It runs at about 10fps on images of 640 x 480 pixels. About 95% of the computational time of each node main computer is devoted to the background subtraction. The appearance models are computed on at least one additional computer at each node. The FRM has not been optimized yet and is called at every 10 sec. by the MONNET system in order to avoid slowing down too much the processing time.

4. CONCLUSION

We have reported about a face recognition module that has been developed for a multi-camera video surveillance system. The module automatically recognizes facial expressions and builds a face model that can be compared to other face models in order to determine if the person has already been seen or not. Six basic facial expressions are taken into account: anger, disgust, fear, happiness, sadness, and surprise, as well as the neutral state. The module is currently being extended and adapted for applications in content-based video indexing and video descriptions[17].

Future improvements of the MONNET system could include (1) elimination of the need for camera color calibration (required by the color appearance model), (2) improving the tracking to better manage pedestrian occlusion and (3) merging the appearance models and send them to the relevant nodes when two appearance models are recognized as representing the same pedestrian (to avoid processing too many models representing the same pedestrian in the node databases).

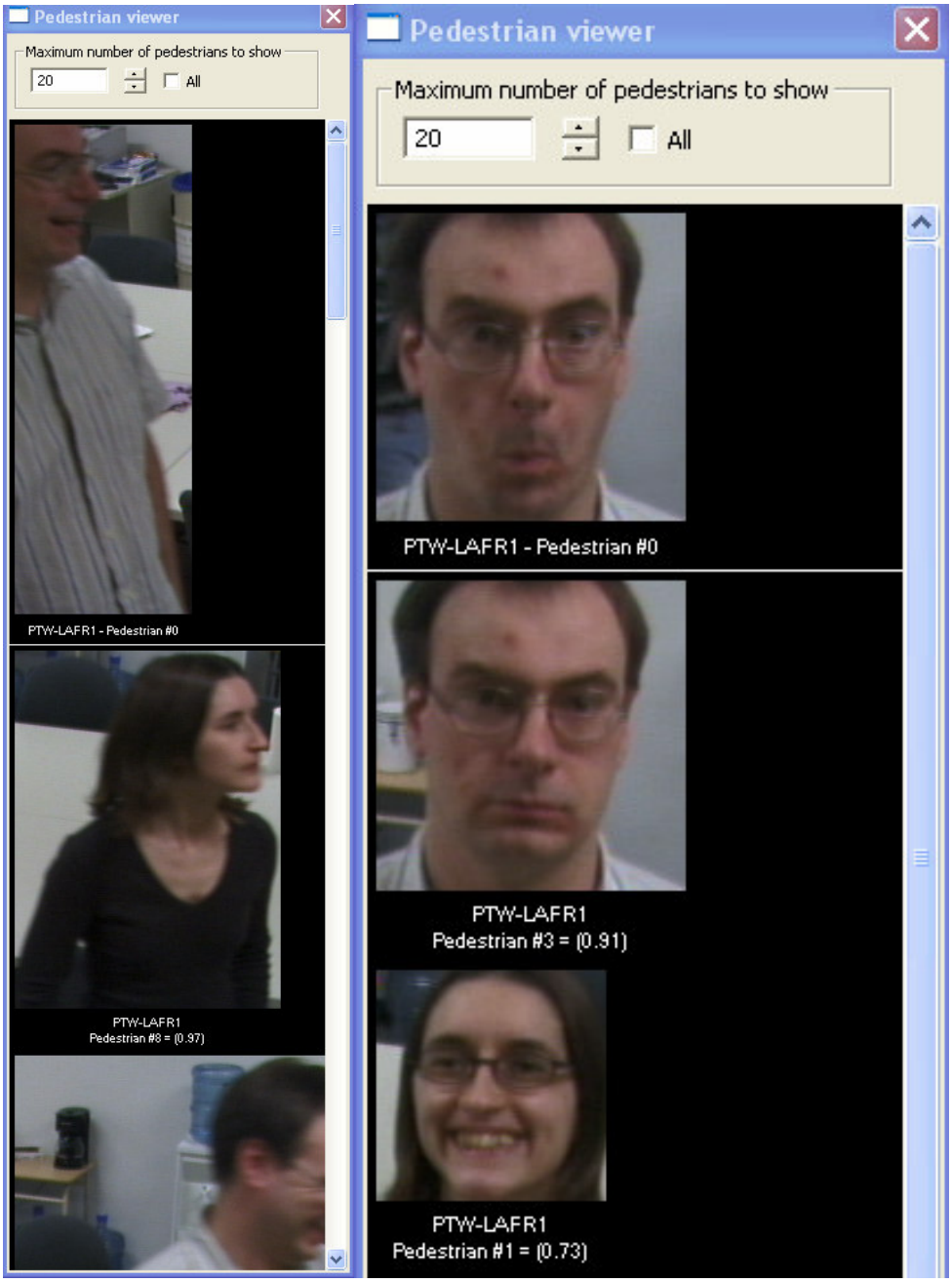

Figure 6: Example where the color appearance model fail and where the face model succeed. Combining multiple source of appearance information makes the system more robust.

ACKNOWLEDGMENTS

This work was supported in part by (1) Precarn Inc., (2) Natural Science and Engineering Research Council (NSERC) of Canada, (3) Defence Research and Development Canada (DRDC) and (4) Ministère du Développement Économique de l'Innovation et de l'Exportation (MDEIE) of Gouvernement du Québec.

REFERENCES

1. D. Laurendeau, A. Branzan Albu, A. Zaccarin, P. Hebert, M. Parizeau, X. Maldague, R. Bergevin, R. Drouin, S. Drouin, N. Martel-Brisson, D. Ouellet, S. Comtois, F. Jean, H. Torresan, F. Laliberté, L. Gagnon, "MONNET: Monitoring Pedestrians with a Network of Loosely-Coupled Cameras", ICPR 2006 (submitted Dec. 2005)
2. R. Collins, A. Lipton, H. Fujiyoshi, T. Kanade, "Algorithms for cooperative multisensor surveillance", Proceedings of IEEE, Vol. 89, No. 10, pp. 1456-1477, 2001
3. C. Stauffer, W.E. Grimson, "Learning patterns of activity using real-time tracking", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 22, No. 8, pp. 747-757, 2000
4. O. Javed, Z. Rasheed, O. Alatas, M. Shah, "MKnight: A Real Time Surveillance System for Multiple Overlapping and Non-Overlapping Cameras", Invited paper in IEEE conf. on Multimedia and Expo, Special Session on Multi-Camera Surveillance Systems, Baltimore, 2003
5. R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, L. Wixson, "A system for video surveillance and monitoring", Technical Report, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, 2000
6. F. Laliberté, "Projet MONNET – Revue de la littérature sur l'analyse des expressions faciales", Technical Report, CRIM, ISBN 2-89522-037-9, 2004 (in French)
7. P. Ekman, W. V. Friesen, "Facial action coding system", Consulting Psychologists Press, Palo Alto, California, USA, 1978
8. F. Laliberté, S. Foucher, L. Gagnon, "A System for Face Characterization of Pedestrians during Monitoring of Extended Premises", Computer Vision and Image Understanding, (submitted Oct. 2005)
9. N. Martel, A. Zaccarin, "Moving Cast Shadow Detection from a Gaussian Mixture Shadow Model", Proc. of IEEE conf. on Computer Vision and Pattern Recognition (CVPR), (San Diego, CA), pp. 643-648, 2005
10. H. Torresan, B. Turgeon, C. Ibarra-Castanedo, P. Hébert, X. Maldague, "Advanced surveillance systems: combining video and thermal imagery for pedestrian detection", Proceedings of SPIE-Defence & Security, Vol. 5405 Thermosense XXVI, pp. 506-515, 2004
11. F. Jean, R. Bergevin, A. Branzan Albu, "Body Tracking in Human Walk from Monocular Video Sequences", in IEEE Canadian Conference on Computer and Robot Vision (CRV), (Victoria, BC), pp. 144-151, 2005
12. M. Lantagne, M. Parizeau, R. Bergevin, "VIP: Vision tool for comparing Images of People", Proceedings of the 16th IEEE Conf. on Vision Interface, pp. 35-42, 2003
13. BioID Face Database: http://www.humanscan.de/support/downloads/facedb.php
14. The FGNET standard: http://www-prima.inrialpes.fr/FGnet/html/home.html
15. Facial annotation database used in this project: http://www.crim.ca/fr/vis-projets.html
16. L. G. Farkas, "Anthropometry of the head and face", Raven Press, 2nd edition, 1994.
17. L. Gagnon, S. Foucher, F. Laliberté, M. Lalonde, M. Beaulieu, "Toward an Application of Content-Based Video Indexing to Computer-Assisted Descriptive Video", CRV 2006 (to appear)