MVA2000 IAPR Workshop on Machine Vision Applications, Nov. 28–30, 2000, The University of Tokyo, Japan

1—3

Partial automation of database acquisition in the FAVRET face tracking and recognition system using a bootstrap approach

Simon CLIPPINGDALE, Takayuki ITO
NHK (Japan Broadcasting Corporation)
Science & Technology Research Laboratories*

Abstract

The FAVRET face recognition system uses a database built from multiple labeled views of each individual to achieve both profile-to-profile tracking of face pose and pose-invariant recognition. Adding a new individual to the database, however, is non-trivial. This work describes the use of a modified version of the FAVRET system as a module in a graphical user interface (GUI) for database acquisition, which automatically detects and tracks face regions in input video or still-frame material and estimates the data which would otherwise have to be entered manually. The estimates can be adjusted manually if necessary. The first application is to update the database in the GUI module itself, using data more accurate than that from which the original prototype database was built.

1. Introduction

The prototype FAVRET face recognition system [1][2] uses a database built from multiple labeled views of about 20 individuals to achieve both pose-invariant recognition and tracking of face pose from profile to profile. The system is being developed with a view to applications in video indexing and scene description, including video editing support.

Currently, the database is built from 19 views of each individual taken at nominally 10-degree intervals from –90 degrees to +90 degrees, where 0 degrees denotes frontal pose. An array of such views for one individual is shown in figure 1. Each view is annotated with the positions of a number of feature points: presently we use the 9 feature points shown. These are placed at locations that possess a degree of 2-dimensional image structure locally (at some resolution), to allow automatic localization and tracking of each feature point. Ambiguous regions such as near contours (where there is much one-dimensional but little two-dimensional structure) and regions that tend to be of changeable appearance such as the hair are avoided.

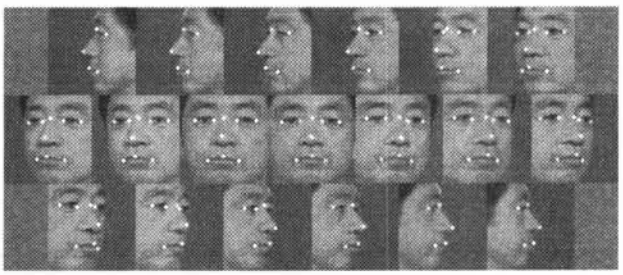

Figure 1 Multiple views at 10-degree pose intervals, annotated with feature point positions.

Rather than the image data itself, the database uses *deformable templates* built from the normalized positions of the feature points and a set of Gabor wavelet features [3][4] computed from the image at multiple resolutions and orientations at each feature point. Each Gabor wavelet feature (coefficient) describes the image intensity in the neighborhood of the feature point and in the neighborhood of a point in the 2-D spatial frequency domain corresponding to the 2-D center frequency of the associated wavelet.

During operation of the FAVRET system, templates from the database are deformed to attain the best fit to face regions in the input video image, producing in the process similarity measures indicating the quality of the match. Both spatial (feature point set distortion) and wavelet-based similarity measures are used for recognition, and the phases of the wavelet coefficients are used to estimate displacement for tracking (see [1][2]). For each candidate face region, a set of hypotheses concerning position, size, pose and identity is maintained and updated in Bayesian fashion according to the results of the deformable template matching at each frame in the input video.

An example of the system output is shown in figure 2. Estimated pose is shown quantized to 10-degree intervals. The "prob" field shows a quantity which resembles (but is not strictly) a posterior probability , to indicate the degree of confidence in the associated estimate of the identity (ID) of the face in question.

*Address: 1-10-11 Kinuta, Setagaya-ku, Tokyo 157-8510 Japan. E-mail: simon@strl.nhk.or.jp

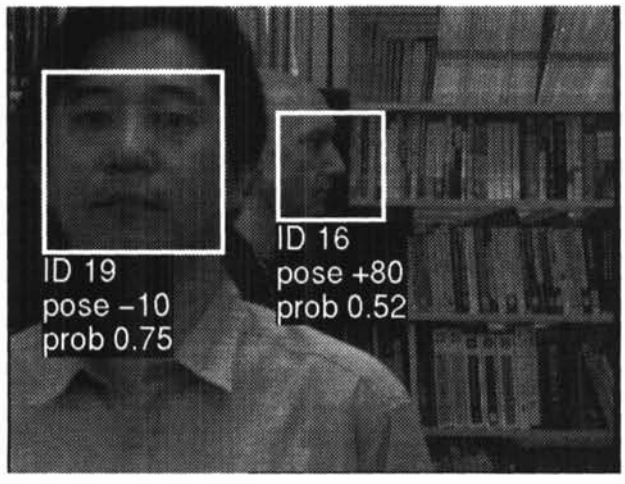

Figure 2 Example of FAVRET system output.

2. Database Registration

Adding a new individual to the database is non-trivial even when multiple views of that individual (such as a video clip) are available. If a video clip is available showing the individual to be registered in poses from profile to profile, the operator is required to perform the following procedure:

1. Select frames showing poses that are as close as possible to multiples of 10 degrees, and do not contain eye blinks, significantly non-neutral facial expressions and so on;
2. On each selected frame, label the feature point positions (as shown in figure 1) with a mouse.

From this data, the database representation (i.e. normalized feature point locations and Gabor wavelet coefficients) is computed.

The registration procedure requires a skilled operator, and even then is attended by a number of problems. These include:

1. *Selection of appropriate frames*: it is difficult to judge head pose visually even to the level of accuracy that the 10-degree interval implies, let alone beyond that. While the recognition performance *per se* of the FAVRET system is not greatly affected by moderate pose errors in the database, the accuracy of the system's pose estimation depends directly upon the accuracy of the poses registered in the database.
2. *Placement of feature points*: there will tend to be systematic differences in placement between different operators, and random variance in placement by any given operator. It is surprisingly difficult to state exactly, to the nearest pixel or two, where a feature such as "the inner corner of the eye" actually is, and this ambiguity is expressed in the variance in the placement of such features by hand.

3. Data Entry GUI

A graphical user interface (GUI) has been developed to assist with data entry. Figure 3 shows the data entry/update window from the GUI.

To assist the operator in the selection of frames showing appropriate poses, the window provides a template in the top half showing a face in the target pose. A frame from the input video, displayed in the lower half, is selected by comparing its pose visually to that of the template.

Once a suitable frame has been selected, feature point locations are entered manually. In an attempt to reduce placement variance, the template is annotated with feature points in the correct positions. The operator may refer to these if unsure of the correct placement.

Which feature point is which is determined entirely by the order in which they are entered, which must be correct: no checks are performed to ensure that the points are in roughly the correct spatial relationship, for example. If two or more feature points are interchanged, performance of the FAVRET system using the new database is likely to suffer.

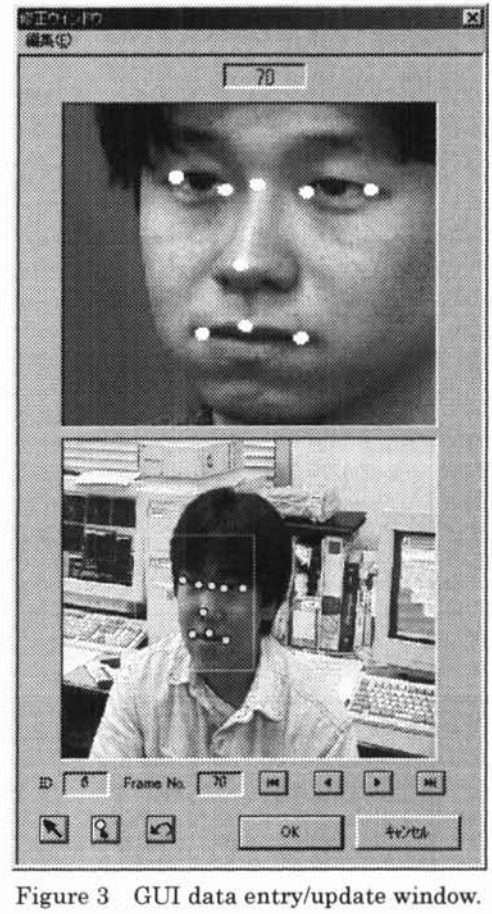

Figure 3 GUI data entry/update window.

4. Partial Automation of GUI Functions

By incorporating into the GUI a module based on the FAVRET system itself, we can partially automate the processes of frame selection and feature point placement. This "GUI FAVRET module" detects and tracks face regions in the input video, producing estimates of face pose and the location of each visible feature point. The feature point estimates are computed from the feature point positions in a number of concurrent hypotheses about face position, pose and identity which the FAVRET system maintains [1][2] for each face region in the input. For the GUI FAVRET module, the identity components of these hypotheses are superfluous since the module does not perform recognition of identity.

Using the pose and feature point estimates produced by its internal FAVRET module, the GUI selects those frames which show poses close to multiples of 10 degrees and offers these frames with their feature point estimates to the user as defaults. The user may adjust the frame selection and/or the feature point locations manually if necessary. Apart from relieving the tedium of performing the entire frame selection and feature point placement operation by hand, this partial automation also removes the danger that the feature points will be entered in the wrong order. The user is presented with feature points already placed (as in the lower half of the window in figure 3) and usually only small adjustments, if any, will be necessary.

Figure 4 illustrates the pose and feature point estimates produced by an experimental version of the GUI FAVRET module.

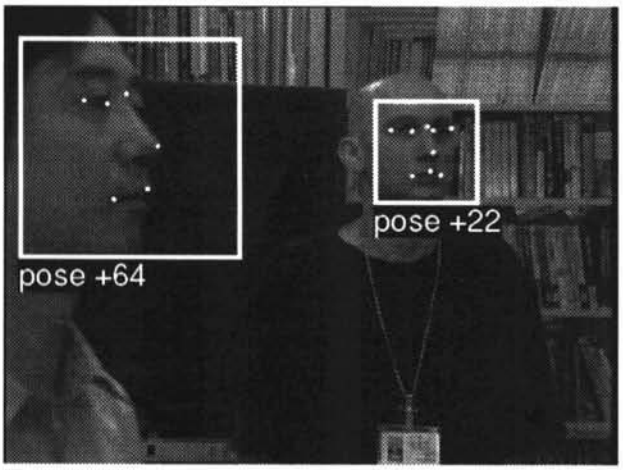

Figure 4 Example output of GUI FAVRET module. Individuals shown were excluded from the module database.

In this example, the module used the original FAVRET system database with the two individuals in the image removed: it is obviously important for

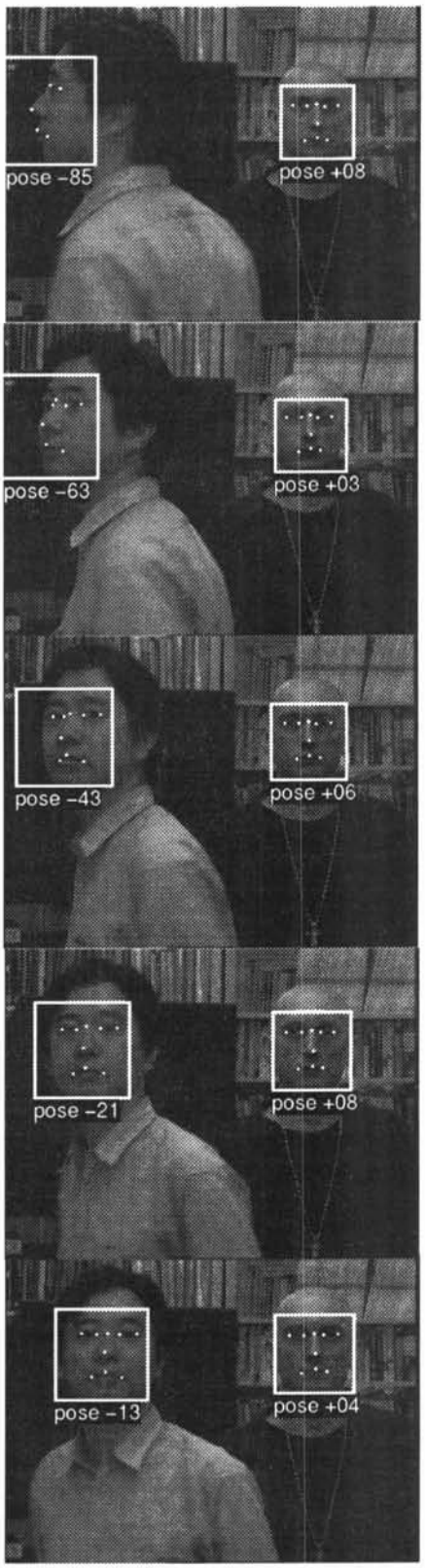

Figure 5 Example output of GUI FAVRET module showing near-profile detection (top) tracked toward frontal. Individuals shown were excluded from the module database.

database acquisition applications that the module be able to handle the case of novel individuals. Experiments suggest that estimates of pose and feature point positions are somewhat noisier for novel individuals than for those registered in the database, but the difference does not appear to be significant.

Where possible, the video clip fed to the GUI and its internal FAVRET module would show just the individual who is to be added to the database. However, there will be situations where the only footage available shows the individual with others. In this case, the module, like the original FAVRET system, tracks all faces which appear in the input video as shown in figure 4. Work is in progress on allowing one individual to be specified as the acquisition target, so that the frames offered to the user will be those in which the target individual appears at the appropriate poses irrespective of the behavior of the other individuals who appear in the same video clip.

As faces in the input turn toward the camera, feature points which were previously occluded become visible and are tracked automatically thereafter. There is no need, for example, to start from a near-frontal pose in which all feature points are visible. The near-profile face in figure 4 has just entered the image from the left and been detected, and the system is correctly tracking those feature points which are currently visible. The near-profile face in the top image in figure 5 has just turned into view from a rear view, and as it turns toward frontal (successive images in figure 5), previously occluded feature points are correctly picked up and tracked.

5. Database Bootstrapping

The database of about 20 persons attached to the prototype FAVRET system (and also used by the prototype GUI FAVRET module) is not very accurate with regard to face pose: the actual pose varies visibly among face images of the same nominal pose. The prototype database also exhibits systematic biases because when the images were taken, the subjects had no reference target marker toward which to turn their heads at each pose. These two factors can lead respectively to dynamic noise and systematic errors in the pose estimates produced by both the FAVRET system and the GUI module based on it. While this may not be a problem in many applications of the FAVRET system itself, it is clearly undesirable that the GUI FAVRET module should propagate pose errors from its own database into the new databases which it is used to create.

For this reason, we have compiled a further set of imagery in which the face pose is much more accurate; a reference marker was provided at eye level with which subjects were required to align their heads. The GUI will be used to label this set (static images are treated as video in which all frames are the same) and then the database used in the GUI FAVRET module itself will be replaced with one built from the new data: hence the term 'bootstrapping.' We expect this to increase the accuracy of the subsequent pose estimation considerably.

6. Further Work

(i) Database acquisition from few views

In the case where we do not have a suitable range of views (such as a profile-to-profile video clip) available, we must attempt to approximate the database representation. This involves estimating feature point positions and Gabor wavelet coefficients corresponding to views at multiples of 10 degrees from just a minimal number of views (two in the worst case, but usually a few more). This is an ongoing research topic ([5]; see also [4]).

(ii) Automatic learning in database

To avoid obsolescence of the database due to people's appearance changing over time, and to replace database representations approximated from few views with measured data when it becomes available in input footage, we hope to incorporate learning functions into the database. These would continually update the database attached to a running FAVRET system from the input video stream when the system is sufficiently confident about the identity of the match.

References

[1] S.Clippingdale, T.Ito, "A Unified Approach to Video Face Detection, Tracking and Recognition", IEICE PRMU 98-200, Osaka, January 1999 (in Japanese).

[2] S.Clippingdale, T.Ito, "A Unified Approach to Video Face Detection, Tracking and Recognition", Proc. ICIP'99, Kobe, Japan, October 1999.

[3] L.Wiskott, J-M Fellous, N.Krüger, C.von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching", TR96-08, Institut für Neuroinformatik, Ruhr-Universität Bochum, 1996.

[4] K.Okada, S.Akamatsu, C.von der Malsburg, "Analysis and Synthesis of Pose Variations of Human Faces by a Linear PCMAP Model and its Application for Pose-Invariant Face Recognition System", Proc. IEEE FG2000, Grenoble, March 2000.

[5] T. Yamane, T. Matsumoto, S.Clippingdale, T.Ito, "Gabor wavelet feature analysis for face pose estimation", Proc. IEICE Annual Conference, Hiroshima, March 2000 (in Japanese).