Video-based framework for face recognition in video

Dmitry O. Gorodnichy
Institute for Information Technology (IIT-ITI)
National Research Council of Canada (NRC-CNRC)
Montreal Rd, M-50, Ottawa, Canada K1A 0R6
http://iit-iti.nrc-cnrc.gc.ca

Abstract

This paper presents a number of new views and techniques claimed to be very important for the problem of face recognition in video (FRiV). First, a clear differentiation is made between photographic facial data and video-acquired facial data as being two different modalities: one providing hard biometrics, the other providing softer biometrics. Second, faces which have the resolution of at least 12 pixels between the eyes are shown to be recognizable by computers just as they are by humans. As a way to deal with low resolution and quality of each individual video frame, the paper offers to use the neuro-associative principle employed by human brain, according to which both memorization and recognition of data are done based on a flow of frames rather than on one frame: synaptic plasticity provides a way to memorize from a sequence, while the collective decision making over time is very suitable for recognition of a sequence. As a benchmark for FRiV approaches, the paper introduces the IIT-NRC video-based database of faces which consists of pairs of low-resolution video clips of unconstrained facial motions. The recognition rate of over 95%, which we achieve on this database, as well as the results obtained on real-time annotation of people on TV allow us to believe that the proposed framework brings us closer to the ultimate benchmark for the FRiV approaches, which is "if you are able to recognize a person, so should the computer".

1 Introduction

The seeming redundancy of the title of this paper, with the first word repeating the last, is attributed to the fact that traditionally approaches to face recognition in video (FRiV) treat video as a collection of images, which are extracted from video and then compared to other images using image-based recognition techniques, of which there are many developed over the long history of face recognition [1]. This paper attempts to challenge this conventional image-based framework to FRiV with another framework, which does not divide video into images, but treats it a whole entity instead. The need for the arrival of such a new video-based framework has been already emphasized in [2, 3] and is also illustrated below.

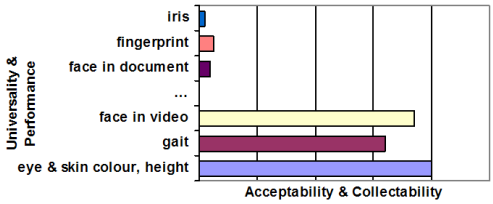

Figure 1: Quality vs availability of different image-based biometric modalities.

1.1 Hard and soft nature of facial biometrics

In the context of biometrics recognition [4, 5], biometric data can be categorized according to their quality and availability as schematically shown in Figure 1, which positions different image-based biometric modalities, according to their quality and availability levels. The extremes on both side of this categorization can be seen: iris recognitionis very robust but very intrusive, person's height or skin colouris not very discriminative but easily collectable and acceptable. Using this figure one can also see that facial data may belong to either side of the biometric modality categorization, as demonstrated below.

ICAO-conformed facial photograph images, which are the images presently used for passport verification and criminal identification and one of which is shown in Figure 2.a, have high resolution (60 pixels between the eyes) and are taken under very controlled lighting conditions (fast exposure, wide aperture and good focus) according to very strict rules, such as: a person has to look straight into the camera, not showing unusual facial expression or wearing any face occluding objects [1]. In this sense, such facial images are not much different from fingerprint images (refer to Figure 1), which also obtained under very controlled conditions, and present the hard biometrics of a person. This type of biometrics, while being very informative becomes

Proceedings of the Second Canadian Conference on Computer and Robot Vision (CRV'05)
0-7695-2319-6/05 $ 20.00 **IEEE**

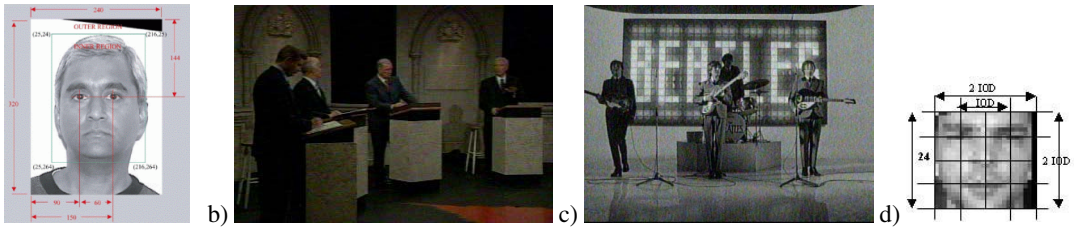

Figure 2: Face image used for face recognition in documents (a), face images obtained from video (b,c), and face model suitable for video-based face processing (d).

available only at a level when a person can be detained and forced to perform certain actions for this biometrics to be measured. Because of this, the number of individuals which can display this type of biometrics is very limited.

On the opposite, facial images acquired from video are very accessible. They can be captured at a distance and can be collected just as easily as eye and skin colour or person's height and weight. They however are much worse in resolution and quality compared to that of photographic facial images. Therefore we can state the following.

Proposition 1: *Facial data as obtained from video and facial data as used for document archival do not belong to the same modality, but rather represent two different biometrics, one being much "softer" than the other.*

What is interesting and paradoxical too is that, while accepting the fact that the video-acquired facial data are of very low quality and resolution, which makes current face recognition approaches very inefficient for face recognition in video [6], a grand challenge for future facial recognition technologies is presently seen (see [1]/FRGC) not in developing new video-based recognition approaches, but in producing high-resolution still images from video. While the latter would indeed improve the performance of FRiV (e.g with techniques from [7, 8]), the objective of this paper is to show that the other way for the improvement of FRiV is also possible and likely even more preferable, on the account of its speed and biological motivation.

1.2 Nominal face resolution

Figures 2.b and 2.c show two video images which are perfectly suited for humans in terms of their ability to recognize the individuals shown in the video. The video sequences, from which the images are taken, are recorded from TV and have the commonly used for TV recordings resolution of 320 by 240 pixels. The faces in these video images occupy 1/16th of the image width. From here we obtain the observation[1] that has become the propelling force for our work.

Proposition 2: *Humans easily recognize a face in video as long as it has resolution of least 12 pixels between the eyes.*

This resolution, which we call *the nominal face resolution* and which is apparently well known to cinematographers, may appear phenomenal for computer scientists working on designing computerized face recognition systems, especially provided that normally in video people do not look into the camera and show quite a lot of orientation and expression variation.

1.3 Biological vision factors

Examining the factors which contribute to the excellent ability of humans to recognize faces in low resolution in video, we can notice the following three factors.

1. We have very *efficient mechanisms to detect a face prior to its recognition*, involving foreground detection and motion/colour tracking, which make recognition easier.
2. Our *decision is based on accumulating results over several frames* rather than on one particular frame and is content dependable, which makes recognition more reliable as we observe a face over a period of time.
3. We use *efficient neuro-associative mechanisms* which allow us a) to accumulate learning data in time by means of adjusting synapses, and b) to associate a visual stimulus to a semantic meaning based on the computed synaptic values.

With the arrival of fast automatic face detectors [9, 10], the first of these factors can be considered practically resolved for video-oriented face recognition systems. The effort on incorporating the second factor into such systems has also been undertaken [11, 12, 13, 14]. This paper, as its predecessors [15, 16], contributes to the effort of other authors working in the field and proposes a computerized version of the third factor. This paper also aims at establishing the common ground for all FRiV approaches by offering a benchmark which can be used to test and refine the approaches.

The organization of the paper is as follows. Section 2 presents a model for the associative processing which is

[1]There does not appear to be a written study regarding this phenomenon. Until such a study arrives, this proposition can be considered as a conjecture, supported by the experiments of this paper and which the readers are encouraged to prove themselves by watching TV, the default resolution of which is 320x240.

Proceedings of the Second Canadian Conference on Computer and Robot Vision (CRV'05)
0-7695-2319-6/05 $ 20.00 **IEEE**

shown to perform well for the problem of accumulation of visual stimuli over time and which is also very efficient in performing the associative recall of the nametags associated with the stimuli. Section 3 presents the video-based facial database compiled in order to provide the benchmark for our framework. Section 4 describes the steps executed within our framework on the way from capturing a video to saying a name of person in it. Section 5 describes the experiments and the neuro-biological statistics used to quantify recognition results and to make time-filtered decisions. Discussions conclude the paper.

2 Modeling associative process

From neuro-biological prospective, memorization and recognition is nothing but two stages of the associative process [17, 18, 19, 20, 21], which can be formalized as follows.

Let us denote an image of a person's face as R (receptor stimulus) and the associated nametag as E (effector stimulus). To associate R to E, let us consider synapses Cij which, for simplicity and because we do not know exactly what is connected in the brain to what, are assumed to interconnect all attributes of stimuli pair R and E among each other. These synapses have to be adjusted in the training stage so that in the recognition stage, when sensing R, which is close to what the system has sensed before, based on the trained synaptic values a sense of the missing corresponding stimulus E is produced.

The following three properties of human brain related to the associative data processing are believed to be of great importance in making strong association: 1) non-linear processing, 2) massively distributed collective decision making, and 3) synaptic plasticity. These properties can be models as follows.

Let $\vec{V} = (Ri, Ei)$ be an aggregated N-dimensional vector made of *all* binary decoded attributes ($R_i, E_i \in \{-1; +1\}$) of the stimuli pair. The synaptic matrix $\mathbf{C}$, which is an NxN matrix, has to be computed so that, when having an incomplete version of a training stimulus $\vec{V}(0)$, the collective decision making results in producing the effector attributes most similar to those used in training. The decision making process is based on summation of all input attributes weighted by the synaptic values, possibly performed several times until the consensus is reached:

$$V_i(t+1) = sign\big(S_j(t)\big), \quad \text{where} \tag{1}$$

$$S_j(t) = \textstyle\sum_{i=1}^{N} CijVj(t), \quad \text{until} \tag{2}$$

$$V_i(t+1) = V_i(t) = V_i(t^*) \tag{3}$$

The last equation, when fulfilled for all neurons, describes the situation of the reached consensus. Thus obtained stimulus $\vec{V}(t^*)$ is called the *attractor* or the *stable state* of the network. It is decoded into receptor and effector components: $\vec{V}(t^*) = (R_i(t^*), E_i(t^*))$ for further analysis of the result of the performed association.

The main question arises: How to compute synaptic values Cij so that the best associative recall is achieved?

Ideally computation of the synaptic values, which is defined by a learning rule, is done so that
i) it does not require the system to go through the already presented stimuli (i.e. there are no iterations involved), and
ii) it would update the synapses based on the currently presented stimuli pair only, without knowing which stimuli will follow (i.e. no batch mode involved).

These two conditions represent the idea of *incremental learning*: starting from zero ($C_{ij}^0 = 0$), indicating that nothing is learnt, each synaptic weight C_{ij} undertakes a small increment dC_{ij}, the value of which, either positive or negative, is determined by the training stimuli pair:

$$C_{ij}^m \quad = C_{ij}^{m-1} + dC_{ij}^m \tag{4}$$

Clearly, the increments dC_{ij}^m should be a function of the current stimulus pair attributes (i.e. $\vec{V}^m$) and what has been previously memorized (i.e. $\mathbf{C}$):

$$dC_{ij}^m = f(\vec{V}^m, \mathbf{C}). \tag{5}$$

The correlation (Hebbian) learning rule, which updates Cij based on the correlation of the corresponding attributes i and j of the training stimulus, is of the form $dC_{ij}^m = f(V_i^m, V_j^m)$, i.e. it makes a default assumption that all training stimuli are equally important as are all attributes i, which is practically never true.

The Widrow-Hoff (delta) rule, which is another commonly used rule:

$$dC_{ij}^m = \alpha V_i^m (V_j^m - S_j^m), \quad 0 < \alpha < 1 \tag{6}$$

is not perfect either, as the learning rate α is the same for all training stimuli, regardless of whether the stimulus is useful or not. This is why this rule is normally used iteratively, applied several times on the entire training sequence until dC_{ij}^m becomes sufficiently close to zero, which makes it unacceptable for applications where training stimuli can not be replayed.

The best incremental learning rule for the binary fully-connected neuron network, in both theoretical and practical sense, is known [22] to be the projection (also called pseudo-inverse) learning rule, which updates the synapses as follows

$$dC_{ij}^m = \tfrac{1}{E^2}(V_i^m - S_i^m)(V_j^m - S_j^m), \quad \text{where} \tag{7}$$

$$E^2 = ||\vec{V} - \mathbf{C}\vec{V}||^2 = N - \textstyle\sum_{i=1}^{N} V_i^m S_i^m \tag{8}$$

is the projective distance which indicates how far the new stimulus is from those already stored.

Proceedings of the Second Canadian Conference on Computer and Robot Vision (CRV'05)
0-7695-2319-6/05 $ 20.00 **IEEE**

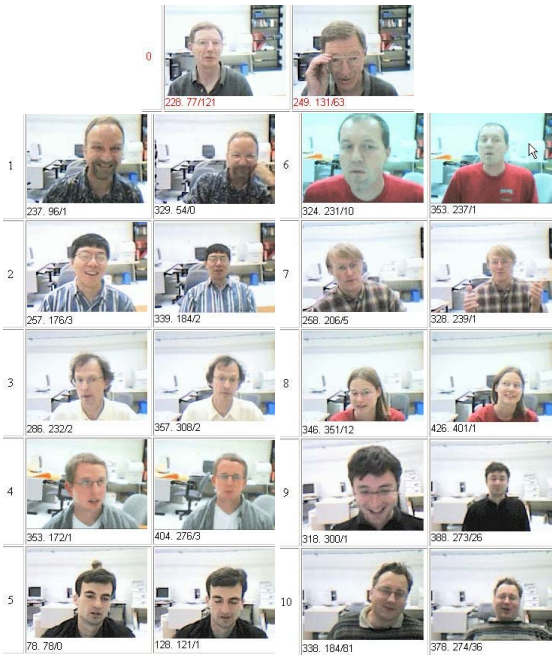

Figure 3: Pairs of 160x120 video clips from the IIT-NRC database. The numbers underneath the images (N.Y/Z) indicate the number of frames in a clip (N) and the number of those of them where one face region (Y) or more (Z), which is the indication of false detection, were detected.

The projection rule is guaranteed to converge to an attractor using the synchronous dynamics [23], which makes the network fast in recognition, and can be further enhanced by reducing the synaptic self-connections as:

$$C_{ii} = D * C_{ii}, \quad 0.05 < D < 0.15 \tag{9}$$

to create a memory of the highest possible capacity and error correction for the given network size [24].

While this model may look too much of a simplification compared to the actual brain, it does cover many properties of the brain [25, 26], such as the binary nature of neuron states, the non-binary nature of inhibitory and excitatory synapses, which are gradually tuned according to the stimulus-response correlation, etc. The assumption of full connectivity allows one to model a highly interconnected network, where the weights of the synapses that do not exist will automatically approach zero as the training progresses.

3 Video-based facial database

In order to provide performance evaluation criteria for the techniques developed and to be developed for face recognition in video and also in order study the effect of different factors and parameters, of which there many influencing the recognition performance in the long chain from processing video to saying a person's name, we have compiled a video-based face database made publicly available at [27].

This database contains pairs of short low-resolution mpeg1-encoded video clips, each showing a face of a computer user sitting in front of the monitor exhibiting a wide range of facial expressions and orientations as captured by a USB webcam mounted on the computer monitor. The driving force for the creation of this database was the goal to examine the computer's ability to recognize faces in conditions known to be sufficient for humans to recognize the faces, in particular in the conditions of low resolution close the nominal face resolution of 12 pixels between the eyes. The video capture size is thus kept to 160 x 120 pixels. With a face occupying 1/4 to 1/8 of the image (in width), this translates into a commonly observed on a TV screen situation when a face of an actor in a TV show occupies 1/8 to 1/16 of the screen.

Each video clip is about 15 seconds long, has capture rate of 20 fps and is compressed with the AVI Intel codec with bit-rate of 481 Kbps. Because of small resolution and compression, thus created video files of person faces are very small (less than 1Mb), which makes them comparable in size with ICAO-conformed high-resolution face images presently used to archive facial images. This fact is worth noting especially provided that in many cases video-based stored faces are more informative than single-image based ones. This also makes our database easily downloadable and thus easier to be used for testing.

The video clips of each person, two of which are taken one after another, are shot under approximately the same illumination conditions (no sunlight, only ceiling light evenly distributed over the room), the same setup and almost the same background, for all persons in the database. This setup allows one to test the recognition performance with respect to such inherent to video-based recognition factors [2] as a) low resolution, b) motion blur, c) out-of focus factor, d) facial expression variation, e) facial orientation variation, f) occlusions without being affected by illumination.

There are eleven individuals registered in this database shown in Figure 5. In our experiments, ten of them are memorized and one is not. All eleven are then used in recognition. Below follows the descriptions of the memorization and recognition processes.

[2] In order to analyze the recognition performance with respect to illumination changes and camera motion, two other databases are being created: one showing the same individuals captured in a different office (with sunlight) and the other showing the same individuals captured by a hand held video camera.

Proceedings of the Second Canadian Conference on Computer and Robot Vision (CRV'05)
0-7695-2319-6/05 $ 20.00 **IEEE**

Figure 4: Facial regions detected in the IIT-NRC database. Note the variation of facial expression and orientation tolerated by the face detector, and also the false "faces".

4 From video input to neuron output

Biological vision systems employ a number of techniques to localize the visual information in a scene prior to its recognition, of which most prominent are fovea-based saliency-driven focusing of attention and accumulation of the captured retinal images over time [15]. What is interesting is that the stimulus captured by eye retina was found [28] to be transmitted to the primary visual cortex of brain, where it is further processed according to the neuro-biological principles described in Section 2, almost without change. This finding made it possible for blind people to "see" by connecting, via electrodes, the output of a video camera directly to the primary visual cortex. This finding also tells us that our neuro-biologically based processing of video can start right on a pixel level of a video frame, with saliency-based localization implemented by means of computer vision techniques.

4.1 Memorizing faces from video

In order to associate a face observed in a video to a nametag, the following basic tasks have to be carried out for each video frame (see also Figure 5).

Task 1. Face-looking regions are detected using a pre-trained face classifier, the one of which, trained on Haar-like binary wavelets, is available from the OpenCV library [29]. Figure 4 shows some facial regions detected on video-clips of our database. As can been seen (see also Figure 3), a face is not detected in every frame. Besides, sometimes more than one face region is detected, i.e. part of a scene is erroneously detected as a face.

Task 2. Colour and motion information of the video is employed to filter out false faces. In particular, faces should have skin colour within certain limits of the skin model, and should have moved within last several frames.

Task 3. The face is cut from the face region and resized to the nominal resolution of 12 pixels between the eyes. In doing this the following preprocessing steps may or may not be performed: a) detection of the facial orientation within the image plane, and b) eye alignment.

Task 4. The receptor stimulus vector $\vec{R}$ of binary feature attributes is obtained from the extracted face. In doing this the following steps are done: a) image is converted to grey-scale, which is known not to affect recognition performance both for humans and computers [30, 15, 1]; b) the canonical eye-centered 24x24 face model described in [15, 16] and shown in Figure 2.d, is used to select the face region to be used in training; c) binarized versions of the selected region and its two gradient images (vertical and horizontal) are used, where binarization is done by comparing the intensity of each pixel $I(i,j)$ to either the average intensity of the entire image I_{ave} (global normalization), or to the average intensity of the 3x3 neighborhood pixels $I_{ave}(i,j,3,3)$ (local, illumination invariant normalization), as

$$I_{binary}(i,j) = sign\big((I(i,j) - I_{ave}(i,j,3,3)\big). \quad (10)$$

It can be noted that the last step is biologically supported and can also be efficiently computed using the local structure transform used in [10]. In addition, if memory and processing time allows, then other encoding schemes describing the pixel interrelationship, such Haar-like wavelets, can also be used to generate binary features.

Task 5. The effector stimulus feature vector $\vec{E}$ is created to encode the name of the person. This is done by creating a binary vector of the size equal to the number of individuals (11 in our case) in which the neuron corresponding to person's name is activated ($E_{i=ID} = +1$), while all other neurons are kept at rest: ($E_{i\neq ID} = -1$). This vector is appended to vector R obtained in the previous task to create the aggregated binary attributes vector $\vec{V} = (\vec{R}, \vec{E})$. Extra ("void") neurons, similar to the hidden layer neurons used in the multi-layered networks, can also be added to possibly improve the network performance.

Task 6. Finally, the obtained aggregated vector $\vec{V}$ is presented to the associative system described in Section 2 for tuning the synapses according to the incremental learning rule of Eqs. 7-9. In doing this, the usefulness of the current frame is analyzed using the distance dissimilarity measure E computed by in Eq. 8. If E is zero or close to zero, then the current visual stimulus is similar to what has been already seen and can therefore be skipped. Other, image processing based techniques can also be used to disregard similar frames from the consideration.

Except for the last of task, which is known to be optimal for the model, all other tasks may require tuning and further investigation for attaining the best recognition performance.

The entire process from capturing image to memorizing a face along with its ID takes about 60 msecs for the network of size N=587 and 150 msecs for N=1739 on Pentium 4 processor. This allows one to memorize faces from video on fly in real time.

Memory-wise, the model is also very efficient. The amount of memory taken by the network of N neurons is N*(N+1)/2*bytes_per_weight. The division over two is thanks to the fact that the weight matrix is symmetric, which is the requirement for the network to converge to an attractor [23].

Proceedings of the Second Canadian Conference on Computer and Robot Vision (CRV'05)
0-7695-2319-6/05 $ 20.00 **IEEE**

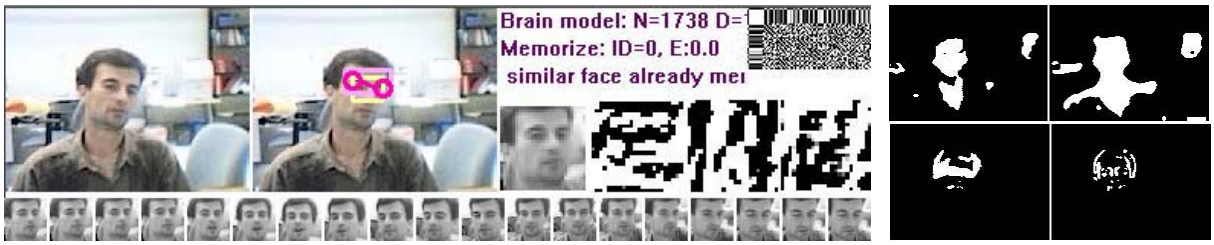

Figure 5: Different stages of memorizing a face from video. When face-looking regions are detected (task 1), they are verified to have skin colour and not to be static inside the white rectangle, using binary colour and change images maps shown at right (task 2). The rotation of the face is detected (15 degrees for the shown frame) using the intensities inside the grey rectangle, and the rotated, eye aligned and resampled to the nominal (12 pixels between the eyes) resolution face is extracted (task 3). The extracted face, shown in the middle, is converted into a binary feature vector (task 4), shown as three binary images. This vector is then appended with the binary representation of the name of the person (task 5) and used to update the synapses of the associative neuron network, the synaptic matrix of which is shown in the top right corner (task 6).

Experiments show that representing weights using one byte (as signed char) is not sufficient, while using two bytes (as float) is. Thus the network of size N=1739, which as shown in next section is quite sufficient for many applications, occupies only 3.5Mb.

4.2 Recognizing faces from video

In recognition, the same chain of steps from video frame capturing to binary encoding of the face image ($\vec{R}^{query}$) is done as in memorization step. The nametag vector $\vec{E}^{query}$, which is appended to the face feature vector, is left unchanged with all neurons at rest: ($\vec{E}_i^{query} = -1$).

The aggregated vector $\vec{V}^{query} = (\vec{R}^{query}, \vec{E}_i^{query})$ is the presented to the network as the initial state $\vec{V}(0)$, starting from which the network evolves according to Eqs. 2-3 until it reaches an attractor. As a result of this association process, one, some or none of the nametag neurons get(s) excited. This neural outcome is analyzed in the context of confidence and repeatability. If several nametag neurons are excited, it means that the system is unsure. At the same time, since the result should be sustainable within short period of time, the same nametag neurons should get excited at least within a few consecutive video frames. Only then a face is considered as recognized. As a possibility and biologically supported, in order to provide a temporal dependence of the current frame from the previous frame, extra neurons can be added to the network to serve as transmitters of the neural outcome obtained on the previous frame to the current frame.

The recognition process from video image capture to telling the person's ID is also very fast. It may seem that, because of many iterations and the large number of neurons, it takes long to compute all postsynaptic potentials S_j in Eq. 2. It is not however, because in every iteration, as proposed in our earlier work [23], instead of considering all neurons for computing S_j as in Eq. 2 we consider only those neurons k, which have changed since the last iteration, to compute S_j as $S_j(t) = S_j(t-1) - 2\sum_k C_{kj} Y_i(t)$ Since the number of these neurons drops down drastically as the network evolves, the number of multiplications becomes very small.

5 Experimental results

The described memorization and recognition process was tested using the database described in Section 3 (and shown in Figure 3). For persons ID=1,...,10, one video clip from the database is used to memorize a face and the other is used for recognition. Person ID=0 is not memorized and is used to test the performance of the system on an unknown person.

5.1 Frame-based recognition

For each video-clip, the following five statistics, derived from the neuro-biological treatment of the recognition process and denoted as S10, S11, S01, S00, and S02, are computed.

S10: The number of frames in a video-clip, in which a face is unambiguisly recognized. These are the cases when only the neuron corresponding to the correct person's ID fired (+1) based on the visual stimulus generated by a frame, the other neurons remaining at rest (-1). This is the best case performance: no hesitation in saying the person's name from a single video frame.

S11: The number of frames, in which a face is not associated with one individual, but rather with several individuals, one of which is the correct one. In this case, the neuron corresponding to the correct person's ID fired (+1), but there were others neurons which fired too. This "hesitating" per-

Proceedings of the Second Canadian Conference on Computer and Robot Vision (CRV'05)
0-7695-2319-6/05 $ 20.00 **IEEE**

formance can also be considered good, as it can be taken into account when making the final decision based on several consecutive video frames. This result can also be used to disregard a frame as "confusing".

S01,S02: The number of frames, in which a face is associated with someone else, i.e. the neuron corresponding to the correct person's ID did not fire (-1), while another nametag neuron corresponding to a different person fired (+1). This is the worst case result. It however is not always bad either. First, when this happens there are often other neurons which fire too, indicating the inconsistent decision – this case is denoted as S02 result. Second, unless this result persists within several consecutive frames (which in most cases it does not) it can also be identified as an invalid result and thus be ignored.

S00: The number of frames, in which a face is not associated with any of the seen faces, i.e. none of the nametag neurons fired. This result can also be considered as a good one, as it indicates that the network does not recognize a person. This is, in fact, what we want the network to produce when it examines a face which has not been previously seen or when it examines a part of the video image which has been erroneously classified as a face by the video processing modules.

Table 1: Frame-based recognition results.

a) Basic case (N=1739):	S10	S11	S01	S 00	S02
ID 1	49	4	0	1	0
ID 2	175	0	3	8	0
ID 3	288	1	2	19	0
ID 4	163	1	11	98	0
ID 5	84	2	3	36	0
ID 6	202	2	3	15	0
ID 7	208	3	12	17	0
ID 8	353	3	8	38	0
ID 9	191	8	30	62	8
ID 10	259	0	10	24	17
Total:	1972	24	82	318	25
ID 0 (unknown face)	0	1	70	112	15
Variations (Totals):					
b) D=1.0	1941	34	46	359	31
c) locally normalized	1821	28	88	555	19
d) intensity only (N=578)	1447	146	258	527	43
e) rotation rectified	1984	20	83	310	24
f) shifted	1964	25	84	321	23
g) added (N=2039)	1971	24	81	325	20
h) trimmed (N=1594)	1562	24	82	318	25
i) hidden (N=1749)	1976	23	81	316	23

The results obtained using the above statistics are given in Table 1. The top part of the table a) shows the results obtained using the basic associative model: the network of N=24*24*3+11=1739 neurons, trained using the projective learning with D=0.1 in Eq. 9 on the binarized canonically eye-centered 24x24 facial image and the two gradient images of it (log file: [27]/log/_111-1739-10.1(7)-10.2(1)-d=0.1.log). The bottom part shows the results obtained using the several variations from the basic model, as mentioned in Section 4.1, namely: b) with change of the learning rule (D=1.0 in Eq. 9), c) using local illumination-invariant binarization (Eq. 10), d) without using gradient images, which results in decreasing the size of the network to N=24*24+11=587, e) with alignment of face rotation prior to recognition, f) with shift of the face area up by one row of pixels, g) with enlarged to 26x26 pixels face area used for feature extraction, which results in increasing the network to N=26*26*3+11=2039; h) without using the boundary and corner pixels of the face model (N=23*23*3-4+11=1594), i) with a few "void" neurons added to increase the memory size and capacity (N=24*24*3+11+10=1749).

Table 2: Neural response in time.

```
Recognition of 05b.avi
*22 -1.0 -0.6 -1.2 -0.7 -0.7 +0.1 -0.5 -1.1 -1.1 -0.7 -1.0
.24 -1.1 -0.6 -1.2 -0.8 -0.8 -0.3 -0.7 -1.3 -1.0 -0.5 -1.3
*26 -1.1 -1.0 -1.0 -0.6 -1.0 +0.2 -0.6 -1.2 -1.1 -0.8 -1.6
...
*70 -1.0 -0.5 -1.1 -0.3 -1.0 +0.4 -0.9 -1.2 -1.3 -1.1 -0.8
+72 -0.8 -0.1 -1.1 +0.2 -1.3 +0.1 -0.6 -0.9 -0.5 -0.9 -0.7
.74 -1.1 -0.5 -1.0 -0.3 -1.3 -0.3 -1.0 -1.0 -1.0 -0.9 -0.8
```

It can be seen that different variations of the associative model do affect the recognition performance of the framework, but not significantly. Most of these observed changes in results can be explained and further analyzed. This however falls outside of the scope of this paper.

5.2 Recognition over time

The data presented in Table 1, while showing the ability of the model to recognize faces from individual low-resolution video frames, do not reflect the dynamical nature of recognition, in particular, the fact that the actual recognition result is based on several consecutive frames rather on each individual frame. Therefore, to understand better the results obtained, the log files showing the neural response in time as well the binaries of our programs are made available at our website. An extract from a log file is shown in Table 2. The rows of numbers in the table show the values of postsynaptic potentials (PSPs) of eleven nametag neurons, carrying the information about the strength of association of a current frame with each of the memorized names, for several consecutive frames (the data are shown for the recognition of person ID=5, every second frame of the video is processed, each line is prefixed with the frame number and *, + or . symbol to indicate the S10, S11 and S00 single-frame outcome).

Proceedings of the Second Canadian Conference on Computer and Robot Vision (CRV'05)
0-7695-2319-6/05 $ 20.00 **IEEE**

Based on these PSPs, the final decision on which nametag neurons "win" and who is the person is made. There are several techniques to make this decision:
a) neural mode: all neurons with PSP greater than a certain threshold $S_j > S0$ are considered as "winning";
b) max mode: the neuron with the maximal PSP wins;
c) time-filtered: average or median of several consecutive frame decisions, each made according to a) or b), is used;
d) PSP time-filtered: technique of a) or b) is used on the averaged (over several consecutive frames) PSPs instead of PSPs of individual frames;
e) any combination of the above.

As mentioned in Section 4.2 and can be seen from Table 2, all of these techniques contribute to a more reliable recognition of faces from video. In particular, they allow one to disregard inconsistent decisions and provide means of detecting frames where a face was falsely or not properly detected by the face detector.

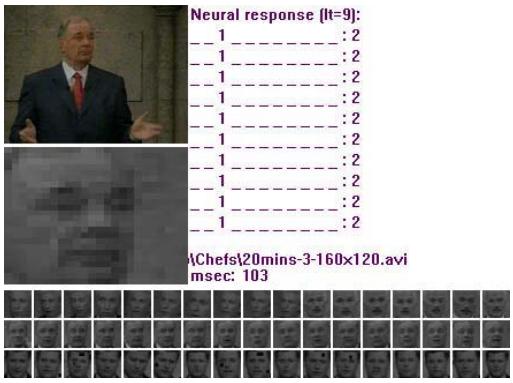

Figure 6: Recognizing a face on TV using the neuro-biological model. For each video frame, the neurons corresponding to the person's ID fire.

6 Discussions

The presented results show that, while some further tuning and testing of the framework may still be required, it does offer a solution to the problem of face recognition in low-resolution videos under unconstrained conditions. In particular, the results achieved are quite sufficient for many applications. One of these applications is designing the perceptual vision systems, such as *N***ouse** [31, 32], which use web-cameras to perceive the commands of computer users and where face recognition can be used to automatically enroll the users so that proper individual settings can be chosen next time they log into the system.

Another application very suitable for the framework and which can also be used as a benchmark for FRiV approaches is automatic annotation of video-conferences and TV programs. To show this, we have recorded a fragment of a TV program, snapshots of which are shown in Figures 2.b and 6, where four invitees (the leaders of Canadian political parties) are debating with each other. To make the test conditions severe, the video is recorded at 160x120 resolution. The face of each invitee was memorized by presenting a 15-secs video-clip showing the face to our program, after which the program is run in recognition mode on the entire prerecorded 20-mins video fragment. The recognition results obtained in this video annotation have been found very promising – practically at every instant, the face of a person was associated with the correct ID . It has to be indicated though that during the entire video fragment, the lighting of the persons did not change, though people were recorded from different view points. This is a situation similar to that of the previous application and that of the IIT-NRC video-based facial database introduced in the paper as a primary testing bench for the framework.

Throughout our we emphasize work that all existing and new video-oriented face recognition technologies should be tested using proper video-based, rather than an image-based, benchmarks. In particular, FRiV approached should be able, after having seen a *video sequence* of a person, to recognize this person in another *video sequence*. The examples of two of such benchmarks are mentioned above. Thanks to the enormous quantity of TV recorded material, the TV-based face recognition testing can be done at different levels of complexity: starting from low-complexity scenarios such as recognizing guests within the same talk show (as in Figure 2.b) to middle-complexity scenarios such as recognizing the same musicians on different stages (as in Figure 2.c) to high-complexity scenarios such as recognizing actors or politicians in different movies over different years. Depending on the real-time and memory constraints and also on the level of sophistication of the recognition approach, the number of individuals memorized from video by an associative memory may have to be limited.

Finally, with respect to biometrics and security applications, taking into account the soft nature of the video-based face recognition biometrics, one may find it more appropriate to use this biometric modality for person classification, rather than for person identification, or use it in a combination with other biometric modalities to improve the overall acceptability and performance of biometrics systems.

References

[1] P. J. Philips, P. Grotherand R. J. Michealsand D. M. Blackburnand E. Tabassi, and J. M. Bone, "Face recognition vendor test 2002 results: Overview and summary," in *http://www.frvt.org*, 2003.

Proceedings of the Second Canadian Conference on Computer and Robot Vision (CRV'05)
0-7695-2319-6/05 $ 20.00 **IEEE**

[2] D.O. Gorodnichy, "Introduction to the first workshop on face processing in video," in *First Workshop on Face Processing in Video (FPiV'04) in CD-ROM Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'04), Washington DC, June 2004*.

[3] D.O. Gorodnichy, "Recognizing faces in video requires approaches different from those developed for face recognition in photographs," in *Proceedings of NATO IST - 044 Workshop on Enhancing Information Systems Security through Biometrics. Ottawa, Ontario, Canada. October 18-20*, 2004.

[4] A. Ross and A. K. Jain, "Information fusion in biometrics," in *Pattern Recognition Letters, Vol. 24, Issue 13, pp. 2115-2125, September 2003*.

[5] A.K. Jain, S. Dass, and K.Nandakumar, "Soft biometric traits for personal recognition system," in *Proceedings of International Conference on Biometric Authentication, LNCS 3072, pp. 731-738, Hong Kong*, 2004.

[6] R. Willing, "Airport anti-terror systems flub tests face-recognition technology fails to flag 'suspects'," in *USA TODAY, September 4, 2003. Available at http://www.usatoday.com/usatonline/20030902/5460651s.htm*.

[7] M. Ben-Ezra, A. Zomet, and S. K. Nayar, "Jitter camera: High resolution video from a low resolution detector," in *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR'04), Washington, D.C., USA, pp. 135-142*, 2004.

[8] G. Dedeoglu, T. Kanade, and J. August, "High-zoom video hallucination by exploiting spatio-temporal regularities," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR '04)*, June 2004, vol. 2, pp. 151 – 158.

[9] G. Shakhnarovich, P. A. Viola, and B. Moghaddam, "A unified learning framework for realtime face detection and classification," in *Int. Conf. on Automatic Face and Gesture Recognition (FG 2002), USA, pp. 10-15*, 2002.

[10] B. Froba and C. Kublbeck, "Face tracking by means of continuous detection," in *First Workshop on Face Processing in Video (FPiV'04), Washington DC, June 2004*.

[11] S. Zhou, V. Krueger, and R.Chellappa, "Probabilistic recognition of human faces from video," *Comput. Vis. Image Underst.*, vol. 91, no. 1-2, pp. 214–245, 2003.

[12] K.-C. Lee, J. Ho, M.-H.Yang, and D. Kriegman, "Video-based faces recognition using probabilistic appearance manifolds," in *Proc 2003 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2003), pp. 313-320, vol. 1, Madison*, 2003.

[13] Aleix M. Martinez and Yongbin Zhang, "From static to video: Face recognition using a probabilistic approach," in *First Workshop on Face Processing in Video (FPiV'04), Washington DC, June 2004*.

[14] O. Arandjelovic and R. Cipolla, "Face recognition from image sets using robust kernel resistor-average distance," in *First Workshop on Face Processing in Video (FPiV'04), Washington DC, June 2004*.

[15] Dmitry O. Gorodnichy, "Facial recognition in video," in *Proc. Int. Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA'03), LNCS 2688, pp. 505-514, Guildford, UK*, 2003.

[16] D.O. Gorodnichy and O.P. Gorodnichy, "Using associative memory principles to enhance perceptual ability of vision systems," in *First Workshop on Face Processing in Video (FPiV'04), Washington DC, June 2004*.

[17] T. Kohonen, "Correlation matrix memories," *IEEE Transactions on Computers*, vol. 21, pp. 353–359, 1972.

[18] W.A. Little, "The existence of the persistent states in the brain," *Mathematical Biosciences*, vol. 19, pp. 101–120, 1974.

[19] S. Amari, "Neural theory of association and concept formation," in *Biological Cybernetics, vol 26, pp. 175-185*, 1977.

[20] C. von der Malsburg, "The correlation theory of brain function," Tech. Rep. 81-2, Dept. Neurobiology, Max-Plank-Insititute for Biophysical Chemistry, Göttingen, Germany, 1981.

[21] D. J. Amit, *Modeling brain function*, Cambridge Univ. Press, 1989.

[22] D.O. Gorodnichy, "Projection learning vs correlation learning: from Pavlov dogs to face recognition," in *AI'05 Workshop on Correlation learning, Victoria, BC, Canada, May 8-12, 2005*.

[23] D.O. Gorodnichy and A.M. Reznik, "Static and dynamic attractors of autoassociative neural networks," in *Proc. Int. Conf. on Image Analysis and Processing (ICIAP'97), Vol. II (LNCS, Vol. 1311), pp. 238-245, Springer*, 1997.

[24] D.O. Gorodnichy, "The optimal value of self-connection or how to attain the best performance with limited size memory," in *Proc. Int. Joint Conf. on Neural Networks IJCNN'99 (Best Presentation award), Washington DC, USA, July*, 1999.

[25] M. Perus, "Visual memory," in *Proc. Info. Soc.'01 / vol. Cogn. Neurosci. (eds. D.B. Vodusek, G. Repovs), pp. 76-79*, 2001.

[26] T. Gisiger, S. Dehaene, and J. Changeux, "Computational models of association cortex," in *Curr. Opin. Neurobiol. 10:250-259*, 2000.

[27] Website, "IIT-NRC facial video database," in http://synapse.vit.iit.nrc.ca/db/video/faces/cvglab.

[28] Wm. H. Dobelle, "Artificial vision for the blind by connecting a television camera to the visual cortex," in *Journal of the American Society for Artificial Internal Organs (ASAIO), 46:3-9*, 2000.

[29] "Opencv library," in *http://sourceforge.net/projects/opencvlibrary*.

[30] Andrew Yip and Pawan Sinha, "Role of color in face recognition," *MIT Tech. Rep. AIM-2001-035 CBCL-212*, 2001.

[31] D.O. Gorodnichy and G. Roth, "Nouse 'Use your nose as a mouse' perceptual vision technology for hands-free games and interfaces," *Image and Video Computing*, vol. 22, no. 12, pp. 931–942, 2004.

[32] Website, "IIT-NRC perceptual vision interface technology," *www.perceptual-vision.com*, 2001.

Proceedings of the Second Canadian Conference on Computer and Robot Vision (CRV'05)
0-7695-2319-6/05 $ 20.00 **IEEE**