Associative Neural Networks as Means for Low-Resolution Video-Based Recognition

Gorodnichy, D.

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

Associative Neural Networks as Means for Low-Resolution Video-Based Recognition *

Gorodnichy, D.
July 2005

* published at International Joint Conference on Neural Networks (IJCNN'05). Montreal, Quebec, Canada. July 31 - August 4, 2005. NRC 48217.

Canada

In Proc. of IJCNN'05, Montreal, Canada, August 1-4, 2005

Associative neural networks as means for low-resolution video-based recognition

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://synapse.vit.iit.nrc.ca

***Abstract*— Techniques developed for recognition of objects in photographs often fail when applied to recognition of the same objects in video. A critical example of such a situation is seen in face recognition, where many technologies are already intensively used for passport verification and where there is no technology that can reliably identify a person from a surveillance video. The reason for this is that video provides images of much lower quality and resolution than that of photographs. Besides, objects in video are normally captured in unconstrained environments, often under poor lighting, in motion and at a distance. This makes memorization of an object from a single video frame unreliable and recognition based on a single video frame very difficult if even possible. This paper introduces a neuro-associative approach to recognition which can both learn and identify an object from low-resolution low-quality video sequences. This approach is derived from a mathematical model of biological visual memory, in which correlation-based projection learning is used to memorize a face from a video sequence and attractor-based association is performed to recognize a face over several video frames. The approach is demonstrated using a video-based facial database and real-time video annotation of TV shows.**

I. Introduction

Figure 1 shows two facial images: a) an image used in passport-based person identification, where a face has resolution of 60 pixels between the eyes and is taken under very controlled conditions, and b) an image taken from a 320x240 video sequence, where faces have barely 12 pixels between the eyes and exhibit a variety of orientations and expressions.

By looking at these figures, one can see that facial images extracted from video are and *will never be* of the same quality and resolution as studio-made photograph pictures. Hence for face recognition not to fail on video sequences, as they do now [1], new *video-based*, rather than image-based, techniques should be developed [2], [3]. An excellent proof that video-based techniques are possible and also an inspiration for these techniques come from biological vision systems, for let us emphasize that the shown image video image is perfectly suited for humans in terms of their ability to recognize people there.

This paper proposes a biologically motivated video-based approach to face recognition. Examining the factors which contribute to the excellent ability of humans to recognize faces in low resolution in video, we emphasize the following three:
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 [4], [5], [6], the first of these factors can be considered practically resolved. The other two still require thorough investigation. Several authors [7], [8], [9] proposed ways to combine frame based decisions over time using the probabilistic framework. This paper proposes another way to do so, by using a neuro-associative framework. In doing this we also present an implementation of the third of the mentioned factors.

The organization of the paper is as follows. Section II presents a model for the associative memorization based on the projection learning of visual stimuli over time. This model allows one to memorize an object based on several observations rather than based on a single image of an object, thus making it possible to learn a face from a low-resolution video sequence. Section III describes video processing steps executed on the way from capturing a video to saying a person name. Section IV describes the ways to evaluate the performance of a neuro-associative face recognition system as well as the ways to integrate the recognition results over time and presents the results obtained by our approach. Conclusions highlight the important applications of the proposed video-based recognition approach.

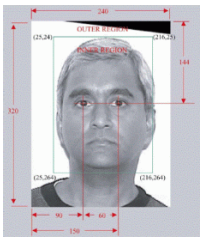
a)  b) 

Fig. 1. Face image used for face recognition in documents (a) and face images obtained from video (b).

II. Modeling associative process

From neuro-biological prospective, memorization and recognition is nothing but two stages of the associative process [10], [11], [12], [13], 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,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 known 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 $\vec{V} = (R_i, E_i)$ be an aggregated N-dimensional vector made of *all* binary decoded attributes ($R_i, E_i \in \{-1; +1\}$) of the stimuli pair. The NxN synaptic matrix $\mathbf{C} = \{C_{ij}\}$ has to be computed so that when having an incomplete version of a training stimulus, the collective decision making produces the effector attributes most similar to those used in training, where the decision making process is based on summation of all input attributes weighted by the synaptic values, performed several times until the consensus is reached:

$$V_i(t+1) = sign\big(S_j(t)\big) \tag{1}$$

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

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

The last equation expresses the *stability condition* of the system. When it holds for all neurons, it describes the situation of the reached consensus. The obtained stimulus $\vec{V}(t^*)$, called the *stable state* or *attractor* of the network, is then decoded into receptor and effector components: $R_i(t^*)$ and $E_i(t^*)$) 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 this should be done so that the computation of the synaptic values defined by a learning rule
i) does not require the system to go through the already presented stimuli, i.e. there are no iterations involved, and
ii) would update the synapses based on the currently presented stimuli pair only, without knowing which stimuli will follow, i.e. no batch mode is involved.
These two conditions represent the idea of *incremental learning*:

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

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.

It is understood that for optimal memorization, the increments dC_{ij}^{m} should be functions of the current stimulus pair attributes (i.e. $\vec{V}^m$) and what has been previously memorized (i.e. $\mathbf{C}$):

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

Unlike the correlation (Hebbian) learning rule of the form

$$dC_{ij}^{m} = \alpha V_i^m V_j^m, \tag{6}$$

which makes a default assumption that all training stimuli and all attributes are equally important, and the Widrow-Hoff (delta) rule of the form

$$dC_{ij}^{m} = \alpha V_i^m (V_j^m - S_j^m), \quad (0 < \alpha < 1) \tag{7}$$

which when applied iteratively on the entire training sequence eventually adjusts synapses to reflect the inter-relationship among the training stimuli, the *projection* (also called *pseudo-inverse*) learning rule, which updates the synapses as

$$dC_{ij}^{m} = \tfrac{1}{D^2(V^m)}(V_i^m - S_i^m)(V_j^m - S_j^m), \quad \text{where} \tag{8}$$

$$D^2(V^m) = ||\vec{V}^m - \mathbf{C}\vec{V}^m||^2 = N - \textstyle\sum_{i=1}^{N} V_i^m S_i^m \tag{9}$$

is both incremental and takes into account the relevance of the training stimuli and their attributes [14], [15]. It is therefore more preferable than the other two for on-fly real-time memorization from video. $D(V^m)$ in Eq. 9 is the projection distance, which indicates how far a new stimulus is from those already stored and which can be used to filter out identical visual stimuli.

The associative model based on the projection learning guarantees convergence of a network to an attractor using synchronous dynamics, as long as the weight matrix is symmetric [16]. This makes the network fast not only in memorization but also in recognition. For this, as proposed in [16] and justified biologically, postsynaptic potentials S_j of Eq. 2 should be computed using only those K neurons , which have changed since the last iteration, as

$$S_j(t) = S_j(t-1) - 2\sum_{i=1}^{K} C_{ij} Y_i(t) \tag{10}$$

Since the number of these neurons drops down drastically as the network evolves, the number of multiplications becomes very small. This makes the model very suitable for memorization and recognition in real time.

Memory-wise, the model is also very efficient. The amount of memory used by the network of N neurons is N(N+1)/2*bytes_per_weight. Experiments show that representing weights using one byte is not sufficient, while using two bytes is [1]. Thus the network of size N=587, which, as will be shown

[1]It should be noted that storing synaptic weights using two bytes leads to breaking several theoretical properties of the pseudo-inverse network, such as $\mathbf{C}^2 = \mathbf{C}$, and as a consequence $|S_i|$ in Eq. 3 may become larger than 1. As our experiments show, this does not produce adverse effects, if properly taken care of.

later, exhibits good associative performance for recognition of up to 10 faces, occupies less than 0.5Mb on hard drive, while the network of size N=1739 occupies only 3.5Mb.

It should also be noted that projection learning allows one to further improve the recognition performance of the system by reducing the synaptic self-connections as:

$$C_{ii} = d * C_{ii}, \quad 0.05 < d < 0.15 \tag{11}$$

to create a memory of the highest possible capacity and error correction for the given network size. This phenomenon, known as the *desaturation* of the network, has been previously shown theoretically [14] and for the case of random patterns in [15]. Subsequent sections also demonstrate it using video-based face recognition.

There are several other ways mentioned in literature for improving the performance of the pseudo-inverse learning using the reduction of the self-connection [17], [18], [19], [20]. Our implementation of those did not show improvement to the recognition performance. The significant slowing down of the memorization and/or recognition process in some cases has been noticed however.

While the presented memorization model may look too much of a simplification compared to the actual brain, it does cover many properties of the brain [21], [22], [23], such as the binary nature of neuron states, the non-binary nature of inhibitory and excitatory synapses tuned according to the stimulus-response correlation, attractor-based dynamics, etc. The thresholds, exceeding which causes physical neurons to fire, are modeled by the self-connection weight values. 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. The study on neurogenesis [24] shows that increasing the neural network size, required to accommodate the increasing number of training stimuli, might also be biologically justified.

What is important is that the described model provides a simple yet efficient means for accumulating knowledge over time, which is what is needed for video-based recognition where each individual video frame, while being of low resolution and quality, cannot be used by itself, but where an accumulation of those can lead to an adequate (i.e. comparable to that of humans) memorization of a face.

III. 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 (e.g. see [25]). What is interesting is that the stimulus captured by eye retina is transmitted to the primary visual cortex of brain, where it is further processed according to the neuro-biological principles described above, almost without a change [26]. 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. It also tells us that associative memorization/recognition of video data can start at a pixel level of a video frame, with saliency-based localization implemented by means of computer vision techniques.

In order to associate a face captured in a video (which serves as receptor stimulus $\vec{R}$ for the associative system) to the person's nametag (which serves as effector stimulus $\vec{E}$) the following chain of tasks is carried out for each video frame (see also Figure 2).

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 [27].

Task 2. Because sometimes parts of a scene are erroneously classified as face regions, colour and motion information of the video is analyzed to filter the spurious face regions.

Task 3. The face is extracted from the face region and resized to the nominal resolution of 12 pixels between the eyes. In doing this, detection of the facial orientation within the image plane and eye alignment are performed.

Task 4. Vector $\vec{R}$ of face attributes is made from the intensities of the extracted face. This is done by using the canonical grey-scale eye-centered 24x24 face model proposed in [25], [28] and shown in Figure 3, which is binarized along with its vertical and horizontal gradient images as follows:

$$I_{binary}(i,j) = sign\big((I(i,j) - I_{ave}\big) \tag{12}$$

where I_{ave} is the average intensity of either the entire image or pixel neighborhood. The latter makes recognition more tolerant to illumination changes, but produces slightly lower recognition rates in illumination constant setups. Other encoding schemes describing the pixel interrelationship, such Haar-like wavelets [5] and local structure transform [6], can also be used to generate binary features, if memory constraints and processing time allow.

Task 5. The effector stimulus feature vector $\vec{E}$, which decodes the face nametag is obtained by fixing the neuron corresponding to the person's ID excited $(+1)$, while keeping other neurons unexcited (-1), with the number of neurons equal to the total number of nametags. When the person's ID is unknown, as in recognition stage, all effector neurons are set unexcited (-1). Extra ("void") neurons, similar to the hidden layer neurons used in multi-layered networks, can be added the network to increase the network capacity and possibly improve the recognition performance. Besides, in order to have a temporal dependency in the recognition process, extra neurons can also be added to the network to serve as transmitters of the neural outcome from the previous frame to the current one.

Task 6. Finally, the obtained aggregated vector $\vec{V} = (\vec{R}, \vec{E})$ is presented to the associative system described in Section II for either performing an association using Eqs. 1-3 (in recognition mode), or tuning the synapses according to the incremental learning rule of Eqs. 8-11 (in memorization mode).

Each of the tasks described offers a variety of research problems and a possibility for improving the system performance. Some of these related to video processing, such as the

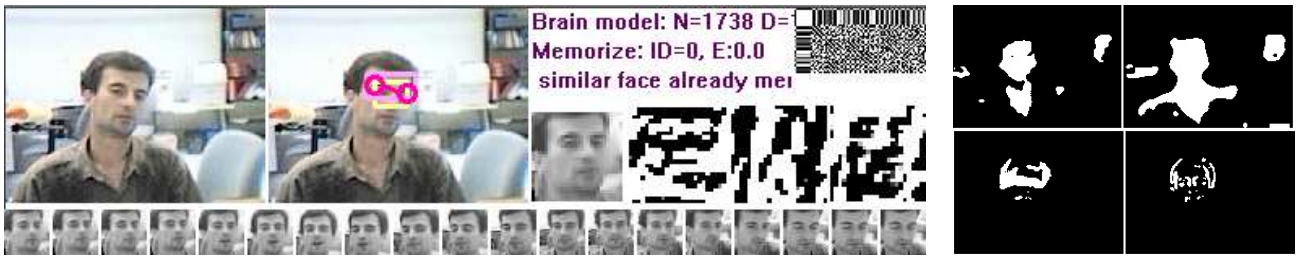


Fig. 2. Different stages of memorizing a face from video. When face-looking regions are detected – task 1, they are verified to have skin colour and not to be static (inside the white rectangle), using binary colour and change images maps (shown at right) – task 2. The rotation of the face is detected (using the intensities inside the grey rectangle) and the rotated, eye aligned and resampled to the 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 (shown as three binary images) – task 4. This vector is then appended by the binary representation of the name of the person – task 5, and is used to update the synapses of the associative neuron network (the synaptic matrix of which is shown in the top right corner) – task 6.

affect of using local illumination-invariant binarization in Eq. 12, alignment of face rotation prior to recognition and using variations to the canonical face model on the performance of the system are addressed in [3]. The others, related to associative neural network model, are studied below.

IV. Nature of neuro-associative recognition

The non-linear neuro-processing, which is performed in the human brain and our recognition system, makes the recognition process different from that of a conventional von-Neumann-type recognition system. In the latter, the decision is normally obtained deterministically based on maximization of an error or probability function. In the former, the decision is based on the binary neural outcome described in terms of the number of the firing nametag neurons This neural outcome has to be 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.

More specifically, the following five frame-based statistics, derived from the neuro-biological treatment of the recognition process and denoted as S10, S11, S01, S00, and S02, are computed for each video fragment.

S10: The number of frames in a fragment, 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" performance 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.

A. Video-based face database

The described approach was tested using the IIT-NRC video-based facial database introduced in [3] and downloadable from [29].This database was created with the goal to examine the computer's ability to recognize faces in conditions known to be sufficient for humans, in particular in the conditions of low resolution close to the nominal face resolution of 12 pixels between the eyes. It contains pairs of 20-second 160x120 mpeg-encoded video clips (see Figure 3), 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 CMOS webcam mounted on the computer monitor. Because of small resolution and compression, video files of person faces in the database are very small (less than 1Mb). This size is comparable to the size of ICAO-conformed high-resolution face images used to archive facial images for forensic purposes, which is worth mentioning, since video is often more informative than a single picture. This also makes the database easily downloadable and thus easier to be used for testing.

Video clips in the database are shot under approximately the same illumination conditions, setup and background. The

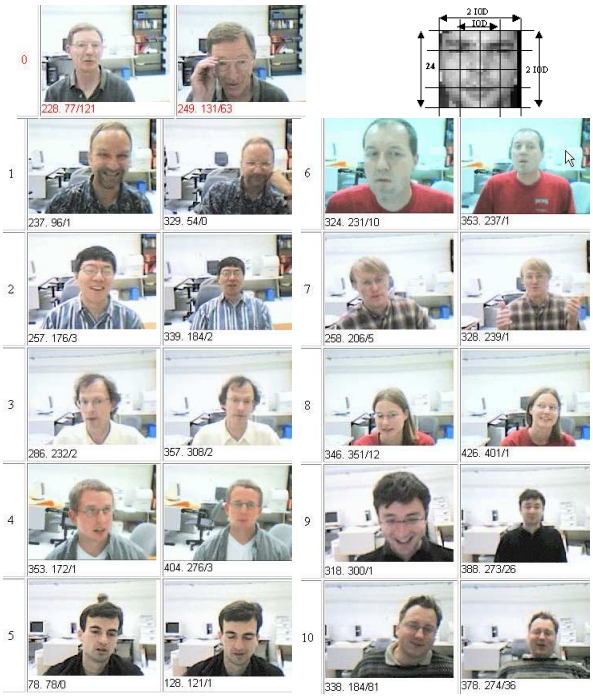


Fig. 3. Pairs of 160x120 video clips from the IIT-NRC database (the numbers underneath the images (N.Y/Z) indicate the number of frames in a clip (N) and the number of those of them where one face region (Y) or more (Z) were detected) and the canonical 12-pixels-between-the-eyes eye-centered face model used for memorizing faces from video.

database is thus most suited for testing the recognition performance with respect to such factors as a) low resolution, b) motion blur, c) out-of focus factor, d) facial expression variation, e) facial orientation variation, and f) occlusion, without taking into account illumination changes.

Table 1 shows frame-based recognition results obtained using our approach for each of eleven persons registered in the database. Ten persons (ID=1,...,10) are memorized, using the first clip of the corresponding video pair, the second clip of the pair is used for testing. One person (ID=0) is not memorized and is thus used to test the performance of the system on an unknown person.

The results are shown for two networks: the 345Kb network of N=24*24+11=587 neurons which uses the intensity values of the image only (left part of the table), and the 3Mb network of N=24*24*3+11=1739 neurons which uses both intensity values and two gradient values of the image (right part of the table). The table also shows total frame-based outcomes for all faces as a function of self-connection reduction coefficient d in Eq. 11.

B. Recognition over time

The data presented in Table 1 show well the ability of the model to recognize faces from individual low-resolution video frames, especially when both face image and its gradient images are used to generate the receptor stimulus. This table however does not reflect the actual nature of neuro-associative

TABLE I

FRAME-BASED RECOGNITION RESULTS

ID	S10	S11	S01	S00	S02	S10	S11	S01	S00	S02
1	48	0	1	5	0	49	4	0	1	0
2	160	7	9	9	1	175	0	3	8	0
3	226	10	18	56	0	288	1	2	19	0
4	78	7	86	96	6	163	1	11	98	0
5	20	2	16	84	3	84	2	3	36	0
6	140	6	22	53	1	202	2	3	15	0
7	187	25	17	10	1	208	3	12	17	0
8	235	60	24	80	3	353	3	8	38	0
9	122	17	42	101	17	191	8	30	62	8
10	231	12	23	33	11	259	0	10	24	17
Total	1447	146	258	527	43	**1972**	**24**	**82**	**318**	**25**
0	0	0	71	110	13	0	0	70	112	15
d										
0.05	1339	198	260	591	43	1975	22	82	317	24
	0	0	76	105	13	0	1	72	113	8
0.10	1447	146	258	527	43	**1972**	**24**	**82**	**318**	**25**
	0	0	71	110	13	0	0	70	112	15
0.15	1459	157	250	514	51	1953	23	87	339	25
	0	0	65	119	10	0	0	69	118	7
1.00	1259	54	111	954	52	1941	34	46	359	31
	0	1	61	135	1	0	0	50	135	9

recognition, which is time-based; in particular, the fact that the final recognition result is based on several consecutive frames rather on each individual frame. Therefore, to see the temporal coherence of the association-based recognition results, the log files of the experiments described above are made available at [29]/log/_100-587-10.1(7)-11.2(1)-d=0.1.log and [29]/log/_111-1739-10.1(7)-10.2(1)-d=0.1.log. An extract from the second of these files is shown in Table 2.

TABLE II

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
```

The rows of numbers in the table show the values of eleven postsynaptic potentials (PSPs) carrying the information about the strength of association of a current frame to each of eleven persons in the IIT-NRC database for several consecutive frames. Each row is prefixed with the frame number and '*', '+' or '.' symbol to indicate the S10, S11 and S00 neural outcome, every second frame of the video being processed. 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 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.

V. Conclusion

The paper described an approach to memorize and recognize faces from low-resolution video using associative neural network. Incremental projection learning is used to accumulate learning data, individual samples of which are of low quality and resolution, over several video frames. Attractor based associative recognition is then used to trace the nametag neurons that fire as a result of associating a new video sequence to the nametags.

The immediate applications of the proposed approach are in i) designing perceptual vision systems such as *N***ouse** [30], which use web-cameras to perceive commands from 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, and ii) video annotation systems, which automatically assign nametags to guests of a TV show [3] or a video-conference [31]. Both of these applications were used as benchmarks for the approach and the results obtained allow us to believe that the proposed approach brings us closer to the ultimate benchmark, which is "if you are able to recognize a person, so should the computer".

Another important application comes from biometrics and security, where soft and unintrusive modality of video-based face recognition can be used in fusion with hard biometric modalities such as fingerprints or studio-taken face photographs to improve the overall acceptability and reliability levels of biometrics systems. It is understood that the same recognition approach can also be used to classify objects other than front faces, for example, head profiles and persons' gaits.

References

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

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

[3] Dmitry O. Gorodnichy, "Video-based framework for face recognition in video," in *Second Workshop on Face Processing in Video (FPiV'05) in Proceedings of Second Canadian Conference on Computer and Robot Vision (CRV'05), 9-11 May 2005, Victoria, BC, Canada*, 2005.

[4] 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.

[5] Rainer Lienhart and Jochen Maydt, "An extended set of haar-like features for rapid object detection," in *IEEE ICIP 2002, Vol. 1, pp. 900-903, Sept.*, 2002.

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

[7] 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.

[8] 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.

[9] 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*.

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

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

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

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

[14] D.O. Gorodnichy and A.M. Reznik, "Increasing attraction of pseudo-inverse autoassociative networks," *Neural Processing Letters*, vol. 5, no. 2, pp. 123–127, 1997.

[15] D.O. Gorodnichy, "The influence of self-connection on the performance of pseudo-inverse autoassociative networks," *Radio Electronics, Computer Science, Control Journal (online at http://csit.narod.ru/journal/riu)*, vol. 2, no. 2, pp. 49–57, 2001.

[16] 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.

[17] H. Ueda, M. Ohta, A. Ogihara, and K. Fukunaga, "An improvement of the pseudoinverse rule with diagonal elements," in *IEICE Trans. Fundamentals, Vol. E77-A, No. 6, pp.1007-1014*, 1994.

[18] S. Ozawa, K. Tsutsumi, and Norio Baba, "A continuous-time model of autoassociative neural memories utilizing the noise-subspace dynamics," in *Neural Processing Letters, Vol. 10, Issue 2, pp. 97-109*, 1999.

[19] N. Davey, R. Adams, and S. P. Hunt, "High performance associative memory models and weight dilution," in *Proc. Int. Conf. on Neural Information Processing ICONIP'01, Vol 2, pp 597-602, Shanghai, China, Nov*, 2001.

[20] F. Clift and T.R. Martinez, "Improved hopfield nets by training with noisy data," in *Proc. Int. Joint Conf. on Neural Networks IJCNN'01, pages 1138–1143*, 2001.

[21] Leslie G. Ungerleider, Susan M. Courtney, and James V. Haxby, "A neural system for human visual working memory," in *Proc Natl Acad Sci U S A. 95(3): 883890*, 1998.

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

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

[24] E.Gould, A.J. Reeves, M.S.A. Graziano, and C.G. Gross, "Neurogenesis in the neocortex of adult primates," in *Science Oct 15 1999: 548-552*, 1999.

[25] 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.

[26] 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.

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

[28] 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*.

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

[30] 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.

[31] M. Fiala, D. Green, and G. Roth, "A panoramic video and acoustic beamforming sensor for videoconferencing," in *Proc. IEEE Int. Workshop on Haptic Audio Visual Environments and their Applications (HAVE'2004). Ottawa, Canada. October 2-3*, 2004.