Using Associative Memory Principles to Enhance Perceptual Ability of Vision System

Gorodnichy, Dimitry; Gorodnichy, O.P.

Proceedings of the First IEEE Computer Vision and Pattern Recognition (CVPR) Workshop on Face Processing in Video, 2004

National Research Council Canada
Institute for Information Technology

Conseil national de recherches Canada
Institut de technologie de l'information

*Using Associative Memory Principles to Enhance Perceptual Ability of Vision System **

Gorodnichy, D., Gorodnichy, O.P.

June 2004

* published in Proceedings of the First IEEE Computer Vision and Pattern Recognition (CVPR) Workshop on Face Processing in Video. Washington, District of Columbia, USA. June 28, 2004. NRC 47139.

Using associative memory principles to enhance perceptual ability of vision systems

Dmitry O. Gorodnichy * and Oleg P. Gorodnichy+
* Institute for Information Technology, National Research Council of Canada,
Montreal Rd, M-50, Ottawa, Canada K1A 0R6
+ Kiev State University of Information and Communication Technologies,
Solomenskaya Str, 7, Kiev, Ukraine, 03110
{ dmitry, alik }@gorodnichy.net
`www.perceptual-vision.com/memory`

Abstract

The so called associative thinking, which humans are known to perform on every day basis, is attributed to the fact that human brain memorizes information using the dynamical system made of interconnected neurons. Retrieval of information in such a system is accomplished in associative sense; starting from an arbitrary state, which might be an encoded representation of a visual image, the brain activity converges to another state, which is stable and which is what the brain remembers. In this paper we explore the possibility of using an associative memory for the purpose of enhancing the interactive capability of perceptual vision systems. By following the biological memory principles, we show how vision systems can be designed to recognize faces, facial gestures and orientations, using low-end video-cameras and little computational power. In doing that we use the public domain associative memory code.

1 Introduction

As noted in the overview of recent advances in computer vision by *The Industrial Physicist* [1], faster computers and easy to install web-cameras have opened the way for new applications of computer vision. One of the most promising is designing *Perceptual Vision Systems*. These are the systems that use video cameras in order to get the information about a computer user or, in other words, in order to *perceive* the user (see Figure 1). The perceived information can be used to enhance the way a user communicates with a computer. Conversely, it can also make computers more intelligent with respect to the user's actions. For example, the position of the user's face can be converted to the position of a mouse pointer; mouth or eye blink motion can be used as a trigger to launch a program; knowing *who* is sitting in front of the monitor may adjust a user-specific configuration, while not detecting anybody can turn the monitor off. This application along with other human-oriented and highly demanded applications of video, such as surveillance and biometrics, requires the recognition of faces in video.

While face recognition problems are well studied for photographic images, it is arguable whether the approaches developed for still imaginary should be applied to video-based recognition, since video information is very different from its static counterpart. On one hand, because of real-time, bandwidth, and environmental constrains, video images are of rather modest resolution and quality, as compared to photo-images. On the other hand, such a seeming deficiency of video is compensated by the abundance of images due to the dynamic nature of video. Video processing and understanding also has many parallels with biological vision, which provides additional insights and solutions to the problem.

Therefore, it is important to develop approaches other than those developed for still imagery which would make use of the advantages of video for face processing. With this goal in mind, in this paper we present a biologically motivated approach for recognizing faces in video based on the associative memory principles.

The organization of the paper is as follows. In Section 2, we reiterate the lessons on visual perception learnt from biological vision, overview main related computer vision results, and present the design of our perceptual vision system. Section 3 describes the attractor-based neural network which mimics the way the human brain memorizes and recognizes, and in Section 4 we show how it can be used for face processing in video. There we also focus our attention on designing a canonical face model suitable for the task and present the results of the experiments. Last section presents the directions for future work and conclusions.

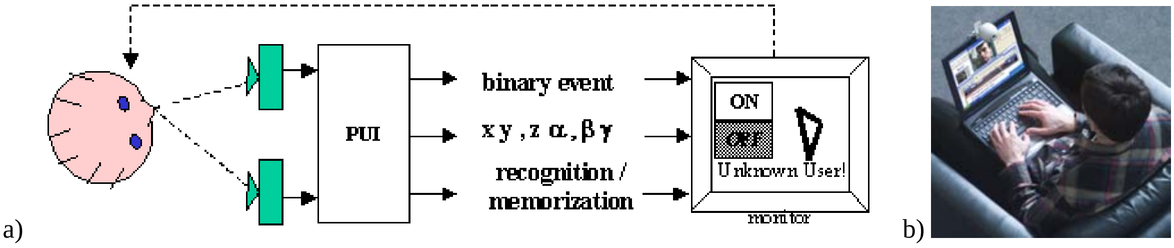


Figure 1: Perceptual vision system: a) diagram and b) setup.

2 From biological to computer vision

The inspiration and foundation for this work comes from both the results obtained by studying biological vision systems and those obtained for computer vision systems. Let us list these results.

2.1 Lessons from biological vision

When recognizing a person, we first scan a scene to localize the areas where the face is, which defines the face segmentation task. In performing this task motion and colour provide the main cues. Then we approach the area of interest and detect the presence of a face there (the face detection task). Then we follow the face (the tracking task), until it appears in the position convenient for recognition, which, in the case of faces, is an eye-to-eye position (eye detection and face modeling tasks). Only then do we attempt to assert whether the face is familiar or not. If it is familiar, we recognize it (the recognition task), and if it does not look familiar, we memorize it (the memorization task).

This is the categorization and hierarchy of face processing tasks, using which we, humans, approach the face recognition problem. A small visual test described in [2] can be used to illustrate this point.

Let us see now how this is explained by the biological vision theory and what are other biological vision mechanisms that make visual recognition efficient in biological systems. Following the work in neurobiology [3, 4, 5] and empirical observations of human visual performance, we find the following results.

1. *Localization vs recognition.* Two separate parts of brain in the visual cortex are responsible for localization ("where") and identification ("what") tasks: in dorsal and ventral steams respectively. This suggests, although this is still debatable (e.g. see [6]), that recognition is performed after an object has been localized. This also suggests that recognition is done in a canonical coordinate space used in the memorization of the object.

2. *Goal-driven vs image-based localization*. When viewing a scene, our eyes do not scan it pixel by pixel, but rather scan it in, what is known as saccadic motion, from one salient point to another. At the same time, a high-level goal, such as "look for faces", also governs the order of scanning a scene. Thus there are two ways for localizing an object: a) top-down, which is driven by a goal, and b) bottom-up, which is driven by the image content. The goal driven deployment of attention however is much slower than the image-based one: 200-1000 ms vs 25-50 ms.

3. *Multi-channel nature of processing.* It has been proposed that factors which contribute to the visual saliency based localization include: a) motion, b) colour, c) intensity gradients (orientation and spatial frequency) and d) disparity (depth). Each component of video is processed very efficiently in retina and the early visual cortical areas, most likely in parallel by different parts of the brain. The winner-take-all neural excitation model is most plausible. This is illustrated by the examples of a frog catching a fly or a bull running on a red cloth of a torero.

4. *Eyes as the most salient features.* Eyes are high contrast moving parts on a face and are therefore very distinctive maximums on saliency maps built in the visual cortex. As such eyes are attended by biological vision systems first. In addition, there are two of them which makes it possible to create the eye mesmorization phenomenon, when one finds himself involuntarily watching straight into the eyes, not being able however too focus on one eye. (The latter is due to the fact that the saliency of a pixel just attended is inhibited to avoid attending it again soon.)

5. *Intensities for recognition.* Study on humans shows that humans make use of colour and motion mainly for the purpose of segmenting a face, rather than for the purpose of recognizing it. At the same time, recognition is performed on the intensity images. However, it is not the intensity values which are used but the gradient, orientation and frequency values of intensity.

6. *Recognition on grey-scale images.* It is also known that humans recognize faces in grey-scale and colour pictures equally well [7]. This became, in fact, such an established phenomenon that even the International Civil Aviation Organization permits using both black-and-white and colour photographs in issuing Machine Readable Travel Documents, such as passports and visas.

7. *Accumulation over time.* The resolution of video in biological vision systems is low everywhere except in the fovea, which is the neighborhood of the point of visual attention (fixation point). Excellent vision-based recognition is achieved therefore by using the already mentioned efficient selective visual attention mechanism and also accumulating the information over time [8]. The discrete way of accumulating video information is most commonly assumed.

2.2 Lessons from computer vision

Besides biological vision results, several good computer-based solutions have been obtained for some face processing tasks. The demonstration of this is the present workshop (see also Table 1). The solutions which we make use of for our work are listed below.

1. *Intensity-based face detection*. The technique by Viola-Jones [9], which uses a pre-trained classifier trained on binary features obtained from face images using Haar-like binary wavelets, allows one to detect close-to-frontal horizontally-aligned faces in still images in close to real time speed. This technique is added to the OpenCV public domain library [10].

2. *Colour for segmentation*. A variety of high-quality human skin colour models have been calculated by several authors for different colour systems. A combination of these models yields “as good as you can get” skin-based face detection in environments with controlled lighting [11]. In environments with unconstrained lighting or with low quality cameras, the recalculation of the skin model is still required however to obtain reasonable face detection. Typical results of skin-based face detection are shown in Figure 2.

3. *Motion for segmentation*. There are well established techniques for background cancellation and foreground detection, based on motion history and accumulation over time. There are also non-linear change detection techniques [12] which are capable of differentiating changes due to the object motion from those due to the lighting changes. They are however quite slow and, for this reason, we do not use them at the moment.

4. *Motion for eye detection*. A recently proposed technique based on computing the second order change, which is the change of the change observed in the image [13], allows one to discriminate the local (most recent) change in image, such as caused by a blink of the eyes, from the global (long lasting) change, such as caused head motion. This allows one to design hands-free blink-operated systems, where “clicking” is performed by *DoubleBlink*, which is a repetitive blinking. This also allows one to precisely localize face from blinking [14].

4. *Intensity gradient nose tracking*. When a face is close enough to capture the convexity of the nose surface, the convex-shape nose tracking technique proposed in [15] allows one to track a liberally unconstrained head motion, represented by the point of the nose tip closest to the camera (or other static object), with sub-pixel accuracy. As shown in [16], this makes it possible to replace such input devices as mouse, track-pad, joysticks, etc. with a hands-free vision-based *Nouse (Nose as mouse)*. Even though it is not as precise as mechanical input devices, it provides a new and very intuitive hands-free way of interacting with a computer. The technique can also be used to verify the position of the face.

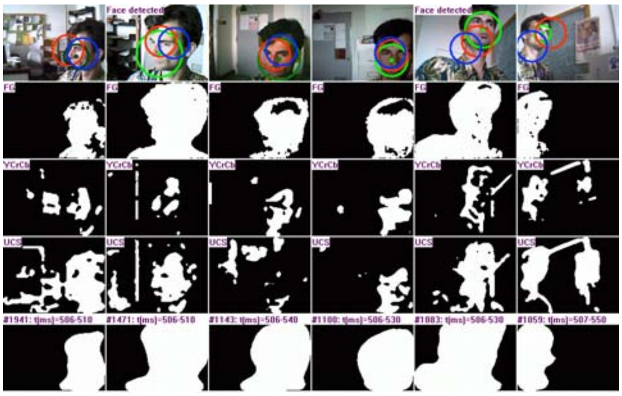


Figure 2: Face detection using colour, motion and intensity components of video with six different webcams. The red and blue circles show the first and the second moments (i.e. location and size) of the binary images which are computed based on colour (for skin) and motion (for change) components of video, respectively. Green circle shows the detection by Viola-Jones face detector followed by the region skin colour based tracking. The binary images are: the foreground images computed using the last 20 video frames and fast dissipation over time (2nd row) and the last 200 video frames using slow dissipation over time (4th row) and the skin images computed using UCS (3rd row) and YCrCb (4th row) colour spaces.

2.3 Building perceptual system

Using the described techniques we build a perceptual vision system capable of detecting, tracking, and localizing faces. The system takes a video sequence as an input, and splits it into the channels corresponding to the motion, colour and intensity components of video. As discussed above, this is how video information is processed by biological vision systems, and this is also, in fact, how face processing tasks are commonly approached by computer scientists. From [2] we see that most computer-vision solutions are single-channel solutions.

The first tasks performed by the system are face segmentation and detection. It is performed by the colour and motion channels which provide the initial estimates for the face size and location. The colour channel uses the combina-

Table 1: Applicability of different face sizes for face processing tasks in 160 x 120 video.

face size in pixels	$\frac{1}{2}$ image 80x80	$\frac{1}{4}$ image 40x40	$\frac{1}{8}$ image 20x20	$\frac{1}{16}$ image 10x10
between eyes	40	20	10	5
eye size	20	10	5	2
nose tip size	10	5	–	–
FS [11, 12]	X	X	X	m
FD [9, 17, 18]	X	X	m	–
FT [19, 20]	X	X	m	–
FL [16, 14]	X	m	–	–
FER [21, 22, 23]	X	X	m	–
FC [9]	X	X	m	–
FM/FI [14, 24, 25]	X	X	–	–

FS, FD, FT, FL, FER, FC, FM/I refer to face segmentation, detection, tracking, localization, expression recognition, classification, and memorization/recognition. (We deliberately make the distinction between FD and FT in that FT uses the past information about the face location, whereas FD does not; and between FT and FL in that FT detects an approximate face location, while FL provides the exact position of a face or facial features. The face size, defined as twice the intra-ocular distance (i.o.d) squared, is given in pixels. X indicates that for a given face size the task can be executed; m signifies that the size is marginally acceptable for the task. The table also shows the anthropometrics of a human face in 160 x 120 video. Seemingly simplified, it, in fact, very well describes the spatial relationship between the facial features (see also Figure 5).

Table 2: Processing time of the perceptual vision system for different resolutions of video (in msecs).

Amount of processing	160x120	320x240	640x480
no processing	10	20	30
colour only	50	80	90
type 1 localization only			
from 1st order change	40	200	400
from 2nd order change	140	600	1400
type 2 localization only	250	320	400

All processing is done on the entire image. Data are given for a reference only, as they depend on the visual content. The second order change detection is performed on the entire image.

tion of the perceptually uniform colour space (UCS) [26] with the nonlinearly transformed YCrCb colour space [27] – these are the two techniques we found to produce the best results – to compute the binarized skin image. The motion channel uses the accumulation over time technique to obtain the binary images of a moving foreground. Analyzing the first and second moments of these binary images yields the initial estimate for size and location of skin coloured and moving regions in video. Figure 2 shows binary skin-map and change-map images obtained using six different webcams (from different manufactures) viewing the same face from different angles. The difference in skin quality detection can be seen.

When a region where a skin-looking moving object is found, the face localization process begins. Following Section 2.1, we consider two types of face localization: Type 1 (visual saliency driven) and Type 2 (goal-driven). Type 1 localization is based on the visual saliency driven deployment of attention. In the context of face recognition, the example of such localization is localizing faces from eye blinking introduced in [14]. Humans have to blink in order to keep their eyes moist. This, as already mentioned, makes eyes the most salient features in a scene. Besides, the eyes blink simultaneously. This provides an additional piece of information which makes eyes and, consequently, the face easy to localize at close range [28].

By Type 2 localization we imply localization of faces driven by the goal of detecting a face. The best example of such localization is the technique of Viola-Jones which works on the intensity channel and which is much slower than Type 1 localization (see Table 2). It however does not need to wait until a person blinks and can therefore localize face more often.

Figure 2 shows typical face detection and localization results obtained by using the three channels of video.

2.3.1 Speed vs resolution vs system complexity

When designing a vision system for perceiving user's facial motions, the question of video processing speed becomes a very important one. Depending on the computer power available, one may wish to sacrifice some face processing in order to maintain the desired video processing rate of at least 10 frames a second needed to capture facial gestures. Table 2 can be used to provide a reference for this question. The data are collected from running the system described above on a laptop, which has 1GHz Celeron processor and 256Mb of RAM. As can be seen from the table, resolution 160x120 is in some cases the only one which permits required real-time processing.

As shown in Table 1, which is obtained by analyzing the related work as well as by conducting our own experiments with perceptual vision systems (downloadable from [29]),this resolution is quite sufficient for many face processing tasks.that for most common setups when a camera is mounted in front of the users face on top of the computer monitor (as in Figure 1), which result in the face occupying more than one quarter of the image, one can safely perform such tasks as nose tracking and eye localization from blinking. This is why we use this resolution in our vision systems.

3 Associative memory

There are a few reasons for the designers of the computer-based recognition systems to look at the way biological memory works. — No doubt, the brain recognizes very fast and very efficiently, compared to any computer-designed memory. When part of the brain is damaged, it can still recognize the data, even from incomplete or changed visual stimulus, which is not always true for a computer memory. In addition, contrary to a computer-designed memory, the brain gets never saturated. That is, it never loses its associative retrieval capability, even when memorizing a practically unlimited number of patterns, even despite the fact that the size of the brain is limited.

Many good qualities of biological memory are attributed to the associative way of memorization performed by the brain.

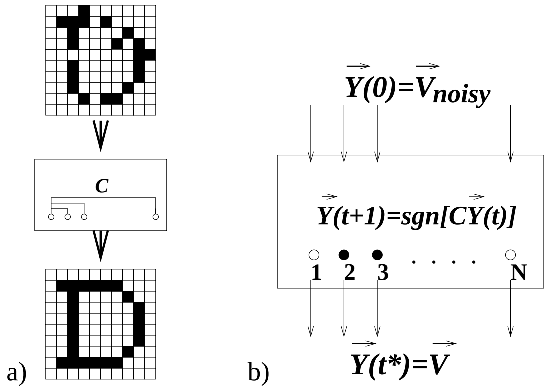


Figure 3: Associative recall of a pattern by an attractor-based neural network.

3.1 Attractor-based model of associative memory

The attempts to mimic associative memory mechanisms of human brain date back to seventies. First, in 1971 and then in 1977 Amari [30] showed that if the formal binary neurons defined in 1943 by McCalloch-Pitts are connected into a fully connected network, then such a system exhibits the associative retrieval. Later W.A. Little [31] showed in 1974 and then in 1978 that indeed the states of the human brain are described by the same equations as the Ising spin-glass neural network, which is a dynamic system equivalent to a fully connected network and that the memories of the brain correspond to the stable states, attractors, of the thus defined neural network. This neural network was found not only to describe well the memorization process happening in the brain, but also easy enough to be evaluated analytically.

Formally described, this network, which we refer to as attractor-based neural network, is made of N fully interconnected dynamic binary active neuron units $Y_i(t), i = 1, ..., N$, which are analogous to the neurons in brain and which can be in either active $(+1)$ or dormant state (-1), depending on the weighted sum of states coming from all other neurons:

$$Y_i(t) = \text{sign}\Big(\sum_{j=1}^{N} C_{ij} Y_j(t-1)\Big). \qquad (1)$$

The information in this neural network is stored in associative sense: starting from an initial unknown state $\vec{Y}(0)$ the network evolves in time according to the update rule Eq. 1 until it reaches an attractor — a stable state $\vec{Y}(t^*)$, which is what the brain remembers and which is given by the stability condition derived from Eq. 1:

$$\vec{Y}^T(t^*)\mathbf{C}\vec{Y}(t^*) > 0, \qquad (2)$$

where $\vec{Y}(t)$ designates a vector made of N values $Y_i(t)$, and $\mathbf{C}$ is an NxN matrix made of synaptic weights C_{ij}.

As seen from Eq. 2, it is the synaptic weights C_{ij} that define the attractors of the memory. The weights also determine how good the associative retrieval capability of the memory is, i.e. how much noise can be tolerated when recognizing an image.

There are many learning rules used to compute synaptic weights C_{ij} for the attractor-based neural networks. However, as we advocate in [32, 33], the best error-correction rate for largest number of prototype patterns stored for these networks is achieved when the *Pseudo-Inverse (PI)* learning rule is used. In particular, this rule is known to be more powerful than other more popular rules such as the *Hebbian inner-product* rule (which is most known as the rule used in Hopfield-like networks and which is, in fact, a simplification of PI rule , made on the assumption that all prototype vectors are orthogonal), and the *Widrow-Hoff delta* iteration rule (which is the commonly used in multi-layered perceptrons and which is another simplification of PI rule which is shown to converge to PI learning rule in large number of iterations).

According to the PI learning rule rule, the weights are computed as a projection matrix, i.e. as the matrix which projects a vector onto the subspace spanned by the prototype vectors:

$$\mathbf{C} = \mathbf{V}\mathbf{V}^+, \qquad (3)$$

where $\mathbf{V}$ is the matrix made of column prototype vectors $\vec{V}^m$, and $\mathbf{V}^+ \doteq (\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T$ is the *pseudo-inverse* of matrix $\mathbf{V}$. This formula is not convenient for implementation because of the matrix inversion involved. Therefore, in simulations the iterative formula is used:

$$\mathbf{C}^m = \mathbf{C}^{m-1} + \frac{(\vec{V}^m - \mathbf{C}^{m-1}\vec{V}^m)(\vec{V}^m - \mathbf{C}^{m-1}\vec{V}^m)^T}{\|\vec{V}^m - \mathbf{C}^{m-1}\vec{V}^m\|^2} \qquad (4)$$

or in scalar form

$$C_{ij}^m = C_{ij}^{m-1} + \frac{(v_i^m - s_i^m)(v_j^m - s_j^m)}{E^2}, \quad \text{where}$$
$$E^2 = N - \sum_{i=1}^N v_i^m s_i^m, \text{ and } s_k^m = \sum_{i=1}^N C_{ik}^{m-1} v_i^m \tag{5}$$

As such, the synaptic weights of the network are symmetric and less the unity in absolute value ($C_{ij} = C_{ji}, |C_{ij}| < 1$).

As shown in [32, 33], by the process called the *desaturation* of the network, which consists in reducing all diagonal weights C_{ii} by some factor (as in Eq. 6 below) in such a way that they never become much larger than the non-diagonal weights C_{ij} , one can memorize with this rule up to $M = 70\%N$ of patterns, achieving the best possible associative performance:

$$C_{ii} := d \cdot C_{ii}, \;\; 0 < d < 0.2. \tag{6}$$

More specifically, the associative retrieval capability of thus built memory (measured by the percentage R of pattern corruption tolerated by the memory) is a function of the number of stored prototypes M, the number of neurons N and the amount the diagonal weight reduction d. See Figure 4 taken from [32, 33] shows this dependency. The figure also shows another interesting results from the theory of these networks which relates the number of states stored by the network M to the ratio of synaptic weights of the network (Cii/Cij). Using this figure one can see, for example, that when the diagonal weights are only 8 times larger than non-diagonal weights, then there are $M = 40\%N$ states stored in the network, and for such number of stores patterns, the network can associatively retrieve them from as much as R=18% of corruption ($N = 500, D = 0.15$ in the figure).

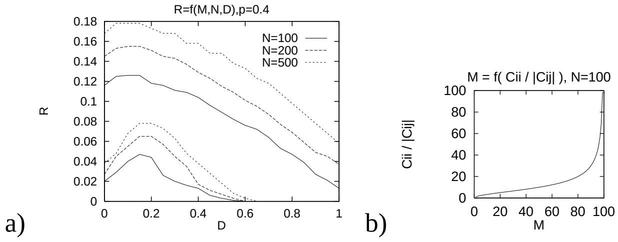


Figure 4: a) the percentage of corruption R from which a memorized pattern is guaranteed to be recovered by the network as a function of network size N, the weight reduction coefficient d and the number of memorized prototypes M ($M = 0.4N$ top three curves, $M = 0.6N$ bottom three curves), and b) The relationship between the number of memorized patterns M and the network synaptic weights (Cii/Cij refers to the average ratio of diagonal and non-diagonal weights).

The desaturation process also allows one to memorize attractors from a continuous flow of prototype data by using the following weight reduction formula:

$$C_{ij}^m = d \cdot C_{ij}^m + (1-d)\frac{(v_i^m - s_i^m)(v_j^m - s_j^m)}{E^2}, \tag{7}$$

with E^2 and s_k^m defined by Eq. 5.

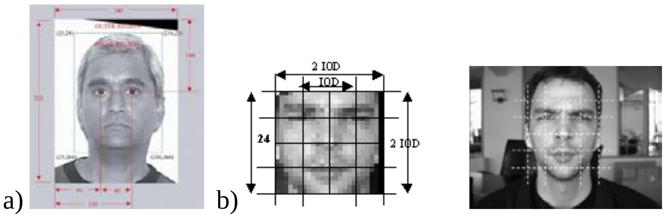


Figure 5: Canonical eye-centered face models: a) as adopted by ICAO for person identification from passport pictures, and b) as designed for recognition of faces in video.

This technique is very useful when memorizing data from continuous stream of video. In this case the network will never saturate, keeping up to $20\%N$ attractors with attractor basins sufficiently large to exhibit associativity. The locations of these attractors in the image space are adjusted automatically as new input prototype vectors arrive.

Finally, it is worth mentioning that the CPP code for designing the described attractor-based memory using the pseudo-inverse learning rule is made publicly available at [34].

4 Using associative memory for facial analysis

By memorizing particular states and then associating the perceived information to one of the memorized states, one can use associative memory to give a meaning to any visual information perceived. There are only two questions needed to be answered before this can be done, namely: How to localize the desired information in the video stream? And how to decode it into the binary form suitable for the neurons of the memory?

In case of face processing, the first question is already answered in Section 2.3. The second one is coupled with two other issues important for face processing, which deal with the canonical face representation and the face encoding scheme.

4.1 Canonical face model

In designing the canonical face representation, which is the representation used for storing the faces, the two key parameters are the location of the face, with respect to the position of the eyes, and its base size, measured by the number of pixels between the eyes.

In [14], in order to facilitate memorization and recognition of faces from blinking, we have designed the eye-centered face representation as shown in Figure 5.b. According to this representation, intra-ocular distance (IOD) defines the area to be used in memorization and recognition. This area is the box with sides being twice the distance between the eyes positioned with respect to the eyes as shown in the figure. While it is fascinating to see that such a box representation divides so neatly a human face into equal blocks, with eyes being at the block intersections, it is the statistical consideration (obtained by examining the BioID database [35]) that has shown that the pixels outside of the box are commonly not very much correlated with the pixels inside the box. The conducted experiments also show that indeed, while being extremely convenient to deal with, this eye-centered four by four block face representation is quite sufficient for many face processing tasks.

For the upper boundary on the face model size, we can refer to the canonical eye-centered face model deployed by the International Civil Aviation Organization (ICAO) for the Machine Readable Travel Documents, which is shown in Figure 5.a. It has 60 pixels between the eyes and is tailored to photograph-based face recognition in databases containing thousands of faces. For perceptual user interfaces however the situation is very different, as it is usually only a few images, such as computer user faces, face orientations, or expressions, which need to be recognized. Hence much smaller face sizes can be chosen for storing the faces.

Our experiments with perceptual vision interfaces show that a person becomes recognizable when there are at least 10 pixels between his/her eyes. This finding is also supported by the study from CMU [36] and MIT [37], which store faces as 20x20 and 19x19 images. This may as well be simply tested by viewing digital video-clips available in abundance in Internet at different resolutions.

Thus, we have decided to use the size of 12 pixels between the eyes for the canonical face model sufficient for perceptual vision systems. Interestingly enough, this 24x24 canonical face representation almost coincides with the base face size used by Viola-Jones; the only difference is that in their model eyes are located one row of pixels lower. Compared to their model, however, our model appears to better capture some facial motions, in particular, mouth opening.

It should also be mentioned that there is yet another reason, besides computational considerations, that for facial recognition in video it is good to represent faces with the least possible size which is sufficient for recognition. This is a well known to the researchers in machine learning the problem of overfitting, which can be formulated, using the Shannon's theory as follows. The data measured with noise should not be modeled with higher precision than what the noise level allows. In the case of video, especially video captured by low-cost cameras and in bad lightning conditions, the corruption of the image is significant. Hence lower resolution is more appropriate.

Table 3: Processing time of performing associative recall on each grabbed frame (in msecs).

face size	24x24	36x36
processing time	100	500

The data are given for the same computer as used in Table 2.

4.2 From pixel intensities to neuron states

As mentioned in Section 2.1 colour does not affect recognition. Therefore, only the intensity values of the face in its canonical representation are used. This results in 24x24=576 eight-bit values representing a face. This can further be processed to reduce the face space. In particular, since in video the eyes can be in both close and open position, the pixels located in the eye positions can be ignored. Also the pixels in the bottom corners can also be ignored.

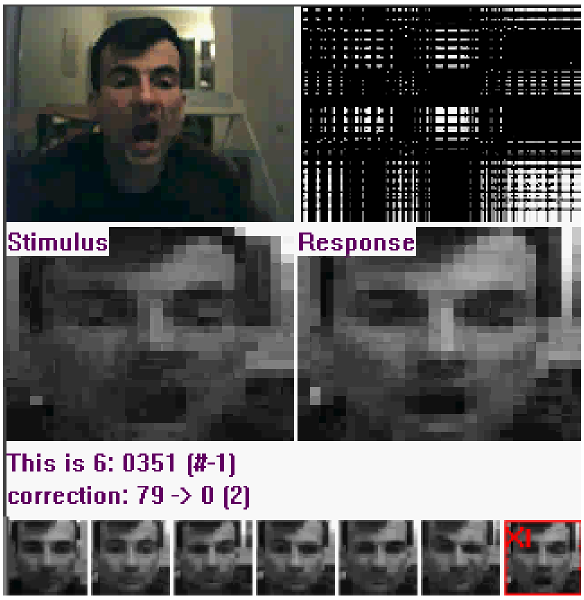
A more biologically justified approach of reducing the face space is to use the spacial redundancy which is present in the face image. This can be achieved by using the Gabor filters, which capture the orientation and frequency components of the image, and which, in this sense, immitate the way the brain processes the image intensities. Gabor transformations however are quite time consuming and therefore we do not use them yet.

Instead, we build an attractor-based memory of size $N = 576$, in which neurons are assigned binarized differences between pixels of the face and the average intensity of the face: $Y_i(0) = sign\big(I_{xy} - I_{ave}\big), i = 1, ..., N$. This size is small enough to keep the system running in real-time (see Table 3). On the other hand, as discussed in the previous section, this size is sufficient to store up to $M = N/2 = 288$ states with the associative retrieval capability of $15\%N$, or up to $M = N/4 = 144$ states with the associative retrieval capability of $25\%N$.

4.3 Experiments

For the purpose of demonstration, we show the results of three experiments in which we apply associative memory to retrieve facial gestures, orientations and identities of computer users.

For the first experiment, we memorized seven facial gestures: normal, both brows up, left brow up, right brow up, right eye wink, left eye wink, mouth open. The result of associative recall is shown in Figure 6. The figure also shows the 576x576 synaptic weight matrix $\mathbf{C}$, in which the value



```
This is 6: 0351 (#-1)
correction: 79 -> 0 (2)
```

Figure 6: Recognizing a facial expression out of 7 memorized.

of weights, are rescaled to the values between 0 and 255, for better viewing. By looking at the image of this matrix and using Figure 4 described in the previous section, one can analytically estimate how much available memory is left and how good the associative retrieval will be (the more noticeable the diagonal of the matrix, the worse the associative property of the memory). In addition, the figure shows the amount of image correction measured as the Hamming distance (79 in this case) between the visual stimulus presented to the network and the associative response of the network, and the number of iterations (2 in this case) taken by network to make the association.

In the second experiment, we memorized eight pan face orientations shown in Figure 7. In this experiment, we also changes the distance from the user to the camera and the lighting conditions, switching the table lamp on and off at different locations. A typical retrieval of face orientation is shown.

Finally, we have memorized all people sitting around the computer, by pointing a camera at each of them and taking three different snapshots of each, after which we continued running the system in the retrieval mode. The representative results are shown in Figure 8.

While the paper shows the snapshots only, the complete video of these and other experiments, the log file with recognition statistics, and the program, with which the results were obtained, are available at our website. The program uses *Microsoft DirectX* [38], *Intel Open CV* [10] and *Intel Image Processing (IPL)* [39] libraries required for image processing and camera functionality and runs on any computer equipped with a USB or firewire camera.

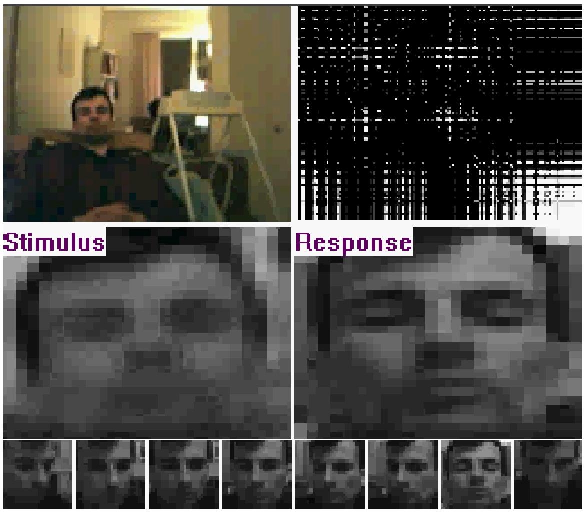


Figure 7: Recognizing a facial orientation out of 8 memorized.

The observations from our experiments can be summarized as follows. The systems works remarkably well, recognizing, in average, seven out of ten frames correctly. The non-linear nature of the network and using binarized intensity differences instead of intensity values of the faces naturally provides a way of dealing with illumination changes, which is a problem for many other face recognition methods.

Although adding the associative recall feature to the perceptual vision system with the makes the system slower, as Table 3 shows (refer also to Table 2), the extra processing time needed to perform an associative recall on a localized face is not significant, being of the same order as the time needed to localize the face. This is one of the advantages of the attractor-based neural network that it memorizes and recognizes very quickly; it takes only one iteration for learning, and three iterations, in average, for recognition. This makes this network very suitable for live video processing applications.

4.4 Temporal filtering

It should be noted now that even when the performance of the system on an individual frame is not very good, it can be drastically improved by applying a temporal filter on the result. As mentioned in Section 2.1, this is what is happening in biological systems. In particular, we presently employ a filter which returns a result as valid only if it is the same within three consecutive video frames. This is a simple trick, which again shows the advantage of video over still imagery.

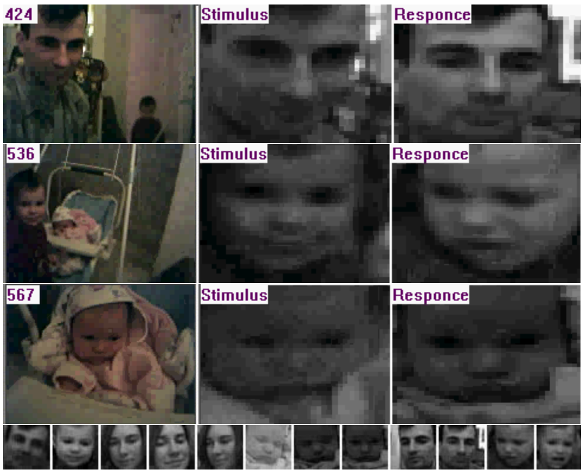


Figure 8: Recognizing users of the system out of 4 memorized: a frame grabbed (left column), a face extracted (middle column), a the association retrieved (third column).

5 Discussions

Up till now associative neural networks are not as commonly used as some other types of neural networks. One may say that this is because there is no advantageous application found for them yet.

As we hope this paper demonstrates, this is now changing. — Using associative neural networks for memorizing and recognizing the perceived video information appears to be just as natural and useful as using other lessons learnt from biological vision for computer vision applications.

This paper serves mainly to introduce the concept of an associative memory, with respect to a new context of analyzing video data, as well as to provide a few examples on how to implement vision-based associative memory using the publicly available CPP code for pseudo-inverse associative neural networks. The examples we considered include recognizing computer users, users' facial gestures and face orientations. We see many promising directions for future work stemming from the results presented. This includes improvements to the presented approaches as well as applying associative memory principles to other vision-based applications.

Better face decoding. As faster computers arrive, one may consider other ways of decoding a face in terms of neuron values, for example, by using binarized differences of all pixels within the face area $Y_i = sign\big(I_{xy} - I_{x'y'}\big)$, Haar-like wavelets as done in [9, 18, 17], or Gabor filters as in [24]. This would increase the size of the neuron network. This however should not be discouraging, should the human brain, which has millions of interconnected neurons, be our inspiration. Also, extra neurons can be added to capture other binary relationships within a face, e.g. based on its shape and geometry.

Several tasks – several networks. The next step would be to employ several networks at the same time, each corresponding to its own task: e.g. one for retrieving a face orientation, one for recognizing a facial gesture, and one for identifying a person. The other option is to use a multilayered structure of networks, with the output of one being the input of another. This idea shows similarity to the actual neuron structure in the brain.

Better generalization. For better generalization, instead of images of one individual, average faces (eigen-faces corresponding to the largest eigen-values) should be memorized (e.g. as in [40]).

Recognizing change and colour states. Finally, being binary in its nature, associative memory is very suitable for dealing with binary images, such as those computed by skin and change detection, some of which are shown in Figure 2. For example, it can be used to associate skin-looking regions to being either a face or non-face. In the same way it can be used to classify certain motion patterns, such as those considered in [28], for example, cause by eye blinking.

Dedication

The authors of the paper dedicate this work to the memory of Ludmila Gorodnichaya, the bright image of whom will always stay in the hearts of her husband, children and grandchildren.

References

[1] M. Picardi and T. Jan, "Recent advances in computer vision," in *The Industrial Physicist, Vol 9, No 1. Online at http://www.aip.org/tip/INPHFA/vol-9/iss-1/p18.html*, 2003.

[2] Dmitry Gorodnichy, "Introduction for the first ieee cvpr workshop on face processing in video," in *Proc. of CVPR Workshop on Face Processing in Video (FPIV'04), Washington DC, 2004*.

[3] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, Mar 2001.

[4] D. Walther, M. Riesenhuber, T. Poggio, L. Itti, and C. Koch, "Towards an integrated model of saliency-based attention and object recognition in the primate's visual system," in *Journal of Cognitive Neuroscience Vol. B14*, 2002.

[5] S.J. Wolfe, J.M. Butcher, C. Lee, and M. Hyle, "Changing your mind: On the contributions of top-down and bottom-up guidance in visual search for feature singletons," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 29, pp. 483–502, 2003.

[6] R. VanRullen and C. Koch, "Is perception discrete or continuous?," *Trends in Cognitive Sciences*, vol. 7, no. 5, pp. 207–213, 2003.

[7] Andrew Yip and Pawan Sinha, "Role of color in face recognition," *MIT tech report (ai.mit.com) AIM-2001-035 CBCL-212*, 2001.

[8] F-F. Li, R. VanRullen, C. Koch, and P. Perona, "Rapid natural scene categorization in the near absence of attention.," *Proc Natl Acad Sci USA*, vol. 99, no. 14, pp. 9596–9601, 2002.

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

[11] J.-C. Terrillon, M. Shirazi, H. Fukamachi, and S. Akamatsu, "Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images," in *Proc. of 4th Int. Conf. on Automatic Face and Gesture Recognition (FG 2000)*, 2000.

[12] E. Durucan and T. Ebrahimi, "Change detection and background extraction by linear algebra," in *IEEE Proc. on Video Communications and Processing for Third Generation Surveillance Systems, 89(10)*, 2001, pp. 1368–1381.

[13] D.O. Gorodnichy, "Second order change detection, and its application to blink-controlled perceptual interfaces," in *Proc. IASTED Conf. on Visualization, Imaging and Image Processing (VIIP 2003), pp. 140-145, Benalmadena, Spain, Sept.8-10*, 2003.

[14] 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.

[15] D.O. Gorodnichy, "On importance of nose for face tracking," in *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG 2002)*, Washington DC, May 20-21 2002, pp. 188–196.

[16] 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. In press, 2004.

[17] Bernhard Froba and Christian Kublbeck, "Face tracking by means of continuous detection," in *Proc. of CVPR Workshop on Face Processing in Video (FPIV'04), Washington DC, 2004*.

[18] Szu-Hao Huang and Shang-Hong Lai, "Detecting faces from color video by using paired wavelet features," in *Proc. of CVPR Workshop on Face Processing in Video (FPIV'04), Washington DC, 2004*.

[19] Le Lu, Xiang tian Dai, and Gregory Hager, "A particle filter without dynamics for robust 3d face tracking," in *Proc. of CVPR Workshop on Face Processing in Video (FPIV'04), Washington DC, 2004*.

[20] Ralph Gross, Iain Matthews, and Simon Baker, "Constructing and fitting active appearance models with occlusion," in *Proc. of CVPR Workshop on Face Processing in Video (FPIV'04), Washington DC, 2004*.

[21] M. Pantic and L. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on PAMI*, vol. 22, no. 12, pp. 1424–1445, 2000.

[22] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier Movellan, "Dynamics of facial expression from video," in *Proc. of CVPR Workshop on Face Processing in Video (FPIV'04), Washington DC, 2004*.

[23] Yingli Tian, "Evaluation of face resolution for expression analysis," in *Proc. of CVPR Workshop on Face Processing in Video (FPIV'04), Washington DC, 2004*.

[24] LeiZhang, Stan Z. Li, and ZhiYiQu, "Boosting local feature based classifiers for face recognition," in *Proc. of CVPR Workshop on Face Processing in Video (FPIV'04), Washington DC, 2004*.

[25] Ognjen Arandjelovic and Roberto Cipolla, "Face recognition from image sets using robust kernel resistor-average distance," in *Proc. of CVPR Workshop on Face Processing in Video (FPIV'04), Washington DC, 2004*.

[26] H. Wu, Q. Chen, and M. Yachida, "Face detection from color images using a fuzzi pattern matching method," *IEEE Transactions on Pattern Anaylis and Machine Intelligence*, vol. 21, no. 6, pp. 557, 1999.

[27] R.-L. Hsu, M. Abdel-Mottaleb, and A. Jain, "Face detection in color images," *IEEE Trans. on Pattern Anaylis and Machine Intelligence*, vol. 24, no. 5, pp. 696, 2002.

[28] Dmitry O. Gorodnichy, "Towards automatic retrieval of blink-based lexicon for persons suffered from brain-stem injury using video cameras," in *Proc. of CVPR Workshop on Face Processing in Video (FPIV'04), Washington DC, 2004*.

[29] Website, "Nouse Perceptual Vision Technology," *http://www.cv.iit.nrc.ca/research/Nouse (www.perceptual-vision.com)*, 2001.

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

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

[32] D.O. Gorodnichy, "The optimal value of self-connection," in *(*Best presentation award *paper, CD-ROM Proc. of Int. Joint Conf. on Neural Networks (IJCNN'99), Washington DC, USA*, 1999.

[33] 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.

[34] Website, "Pseudo-inverse associative memory," http://www.cv.iit.nrc.ca/∼dmitry/pinn, 2000.

[35] The BioID, "Face database," *http://www.bioid.com/ downloads/facedb/facedatabase.html*, 2001.

[36] H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Transactions on PAMI*, vol. 20, no. 1, pp. 23–38, 1998.

[37] Heisele, T. Poggio, and Pontil, "Face detection in still gray images," *MIT tech report (ai.mit.com)*, 2000.

[38] "Microsoft DirectX 8 SDK," in *http://www.microsoft.com/downloads*.

[39] "Intel image processing library (ipl)," in *http://developer.intel.com/software/products/perflib/ijl*.

[40] Mario Romero and Aaron Bobick, "Tracking head yaw by interpolation of template responses," in *Proc. of CVPR Workshop on Face Processing in Video (FPIV'04), Washington DC, 2004*.