

Associative tracking and recognition in video

Dmitry O. Gorodnichy
Computational Video Group
Institute for Information Technology
National Research Council Canada
Montreal Rd, M-50, Ottawa, Canada K1A 0R6
<http://synapse.vit.iit.nrc.ca>

Abstract: Due to limited resolution and quality of surveillance video, tracking and recognizing of objects in surveillance video requires techniques for accumulation of information about the object over time. The simplest of these techniques is histograms, which computes the distribution of pixel values over time and which is frequently used for tracking uniformly coloured objects such as faces. Another well-known technique for the purpose is correlograms, which learns pixel values and their spatial relationship to yield better discriminative power. However, this technique also lacks a comprehensive learning ability, because it updates the information about the object, expressed in terms of cooccurrence matrices, using the currently observed pixels only and ignoring the preceding learning history. Associative neural network based memorization can thus be seen as a more sophisticated mechanism for accumulation of information about an object in a video sequence. This presentation describes how to use the **Open Source Associative Neural Network code** for tracking and recognition of objects in video. Two demos showing multiple-face tracking and classification in low-resolution video are shown.



Figure 1. Learning an object (face) in video over time using associative neural network: binary features from the localized over several frames object are fed into the network along with the binary object ID. As a result, the network readjusts its synapses based on the past learnt history and the current stimulus so that, when the next time an image of an object is shown, the associated ID (or IDs) is/are retrieved.



Associative tracking and recognition in video

Dmitry O. Gorodnichy

Institute for Information Technology,
National Research Council Canada
<http://synapse.vit.iit.nrc.ca/memory>

The First International Workshop on *Video Processing for Security*
June 7-9, Quebec City, Canada



National Research
Council Canada

Conseil national
de recherches Canada

Canada

Proper treatment of video-based tasks

ICAO-conformed passport photograph
(presently used for forensic identification)

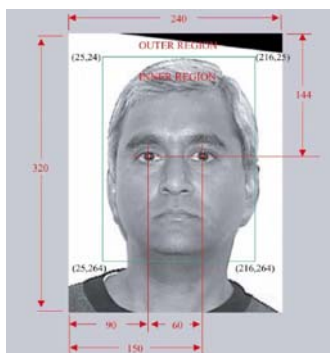
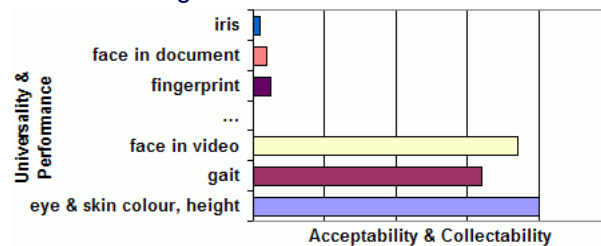


Image-based biometrics modalities



Photographic facial data and video-based facial data are two different modalities:

- different nature of data
- different biometrics
- different approaches
- different applications
- different testing benchmarks

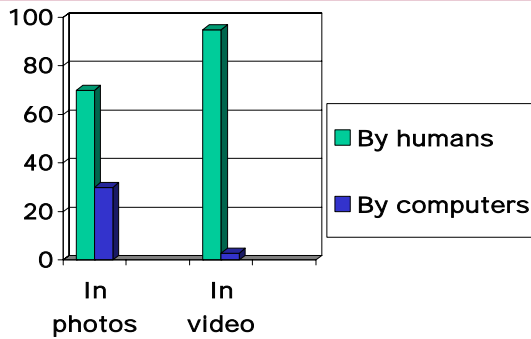
❖ In video: faces are meant to be of low quality and resolution (~ 10-20 pixels IOD)



Images from surveillance cameras (of 11/9 hijackers) and TV.

NB: VCD is 320x240 pixels

Proper tasks for Video Recognition



Face recognition systems performance
(from NATO Biometrics workshop, Ottawa, Oct. 2004)

Goal:

- NOT “in making video data of better quality” (as seen by Face Recognition Grand Challenge – www.frvt.com),
- BUT in **finding approaches & applications most suited for low-quality video-based data**

Good applications of video recognition (esp. wrt faces)

1. Multi-object tracking (in particular, multi-face tracking)
 - a. Back-tracking,
 - b. Multi-camera tracking: simultaneous and in sequence

ASSOCIATIVE TRACKING

2. Recognize what has been already seen
 - limited class number classification / Video annotation

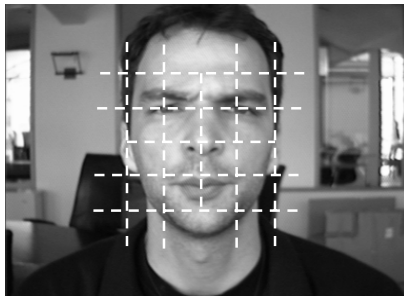
ASSOCIATIVE RECOGNITION

3

Applicability of 160x120 video

According to face anthropometrics

- studied on BioID database
- tested with Perceptual Vision interfaces
- observed in cinematography



Main conclusions for video-surveillance-based biometrics:

1. If $IOD < 10$, no face-based biometrics is possible. Use body biometrics !
2. If $IOD > 10$, face-based recognition is possible, but mainly for limited #faces !

Face size	1/2 image	1/4 image	1/8 image	1/16 image
In pixels	80x80	40x40	20x20	10x10
Between eyes- IOD	40	20	10	5
Eye size	20	10	5	2
Nose size	10	5	-	-
FS	√	√	√	b
FD	√	√	b	-
FT	√	√	b	-
FL	√	b	-	-
FER	√	√	b	-
FC	√	√	b	-
FM / FI	√	√	-	-

FS,FD,FT,FL,FER,FC,FM/I = face segmentation, detection, tracking, localization, expression recognition, classification, memorization/identification

√- good
b - barely applicable
- - not good

4

Photos vs Video

Photos:

- High spatial resolution
- No temporal knowledge

E.g. faces:

- in controlled environment
(*similar to fingerprint registration*)
- “nicely” forced-positioned
- 60 pixels IOD

Video:

- Low spatial resolution
 - High temporal resolution
- (Individual frames of poor quality)

Taken in unconstrained environment (in a “hidden” camera setup)

- don’t look into camera
- even don’t face camera
- 10-20 pixels IOD (intra-ocular distance)

Yet (for humans), video (even of low quality) is often much more informative and useful than a photograph !

Video processing implies accumulation over time:
in recognition and esp. in learning!

5

Why associative memorization?

Techniques to accumulate knowledge (i.e. learn data) over time:

- Histograms – simple count of the same values (in 1D, 2D)
- Next level: Correlograms (geometric histogram) – count of pixels and their inter-relationships
- NEXT level: Associative memorization – not just a count of pixel values and pair-wise pixel relationships, but also takes in account the entire picture of the data: PAST DATA and ALL PRESENT DATA

All of these learn data in real-time

All of these recognize data in real-time

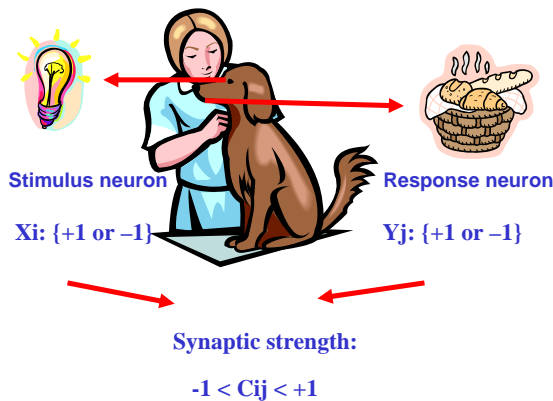
6

From visual image → to saying name

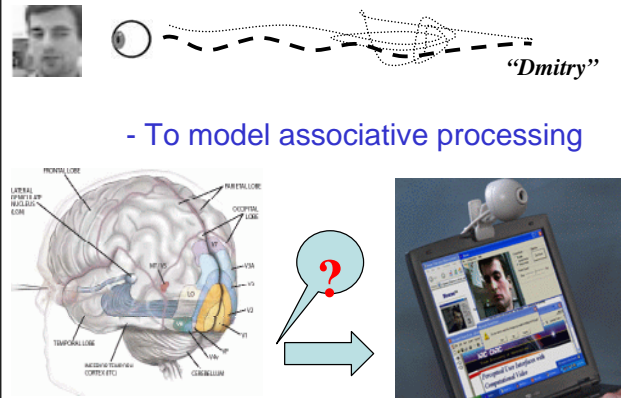
From neuro-biological prospective, memorization and recognition are two stages of the associative process:

From receptor stimulus R → to effector stimulus E

Main associative principle



What do we want ?



Main question of learning: **How to update synaptic weights C_{ij} as $f(X,Y)$?**

Keys to resolving association problem

To understand how human brain does it

I. 12 pixels between the eyes is sufficient !

II. Main three features of human vision recognition system:

1. Efficient visual attention mechanisms
2. Accumulation of data in time
3. **Efficient neuro-associative mechanisms**

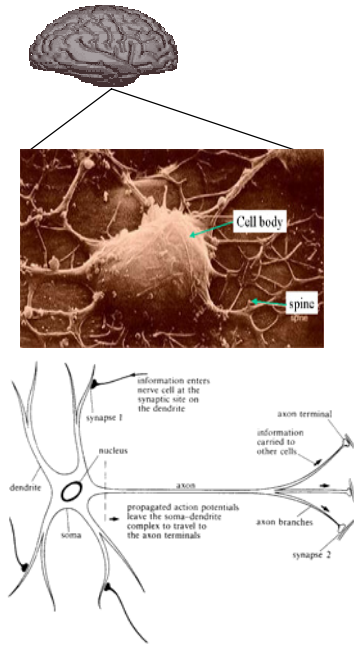
III. Main three neuro-associative principles:

- 3.1 Non-linear processing
- 3.2 Massively distributed collective decision making
- 3.3 Synaptic plasticity

for:

- a) accumulation of learning data in time by adjusting synapses
- b) association a visual stimulus to a semantic meaning based on the computed synaptic values

Lessons from biological memory



- Brain stores information using synapses connecting the neurons.
- In brain: 10^{10} to 10^{13} interconnected neurons
- Neurons are either in rest or activated, depending on values of other neurons Y_j and the strength of synaptic connections: $Y_i = \{+1, -1\}$
- Brain is a network of “binary” neurons evolving in time from initial state (e.g. stimulus coming from retina) until it reaches a stable state – attractor.

$$Y_i(t+1) = \text{sign}\left(\sum_{j=1}^N C_{ij} Y_j(t)\right)$$

- Attractors are our memories!

9

Refs: Hebb'49, Little'74,'78, Willshaw'71

Learning process

Learning rules: From biologically plausible to mathematically justifiable

$$C_{ij}^m = C_{ij}^{m-1} + \Delta C_{ij}^m$$

- Hebb (correlation learning): $\Delta C_{ij}^m = \frac{1}{N} V_i^m V_j^m$ is of form $\Delta C_{ij}^m = \alpha F(V_i^m, V_j^m)$
- Better however is of form: $\Delta C_{ij}^m = \alpha F(C_{ij}^{m-1}, V_i^m, V_j^m)$
- Should be of form: $\Delta C_{ij}^m = \alpha F(C^{m-1}, V^m)$
- Widrow-Hoff's (delta) rule: $C^{k+1} = C^k + \alpha(\vec{V} - C^k \vec{V}) \vec{V}^T$

- **We use Projection Learning rule:**

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

$$E^2 = \|\vec{V}^m - C^{m-1} \vec{V}^m\|^2$$

It is most preferable, as it is:

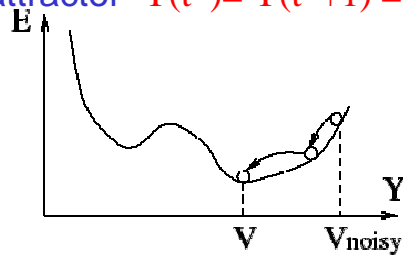
- both incremental and takes into account relevance of training stimuli and attributes;
- guaranteed to converge (obtained from stability condition $\mathbf{V}^m = \mathbf{C} \mathbf{V}^m$);
- fast in both memorization and recognition; also called *pseudo-inverse rule*: $\mathbf{C} = \mathbf{V} \mathbf{V}^+$

Refs: Amari'71,'77, Kohonen'72, Personnaz'85, Kanter-Sompolinsky'86, Gorodnichy'95-'99

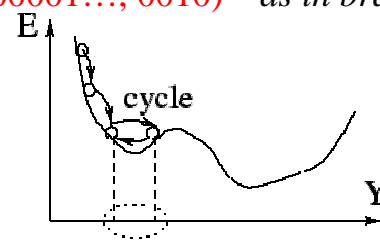
10

Recognition process

- Each frame initializes the system to state $Y(0) = (01000011..., 0000)$ from which associative recall is achieved as a result of convergence to an attractor $Y(t^*) = Y(t^*+1) = (01000001..., 0010)$ – as in brain...



Static attractor



Dynamic attractor

- Effector component of attractor (0010) is analyzed. Possible outcomes: S00 (none of nametag neurons fire), S10 (one fires) and S11 (several fire)
- Final decision is made over several frames:

```
0000100000
0000000000
0000100000
0000100000
```

```
0000100000
0010100000
0000100000
0000100000
```

```
0000100000
0000010000
0000100000
0000100000
0000100000
```

11

(e.g. this is ID=5 in all these cases)

From video input to neural output

- face-looking regions are detected using rapid classifiers.
- they are verified to have skin colour and not to be static.
- face rotation is detected and rotated, eye aligned and resampled to 12-pixels-between-the-eyes resolution face is extracted.
- extracted face is converted to a binary feature vector (Receptor)
- this vector is then appended by nametag vector (Effector)
- synapses of the associative neuron network are updated

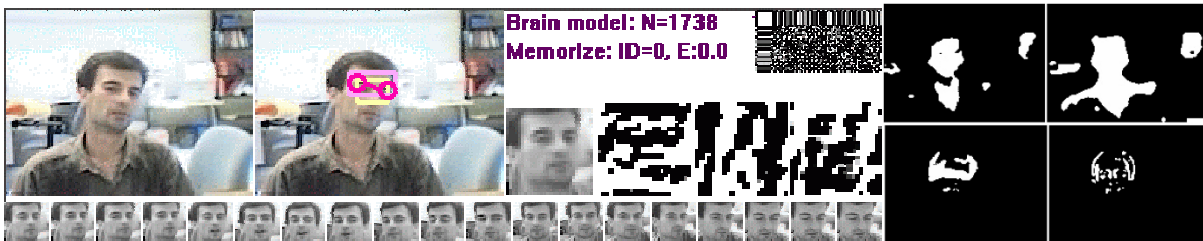


Table 2: Neural response in time.

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
```

Time weighted decision:

- neural mode: all neurons with PSP greater than a certain threshold $S_j > S_0$ are considered as "winning";
- max mode: the neuron with the maximal PSP wins;
- time-filtered: average or median of several consecutive frame decisions, each made according to a) or b), is used;
- PSP time-filtered: technique of a) or b) is used on the averaged (over several consecutive frames) PSPs instead of PSPs of individual frames;
- any combination of the above.

S10 - The numbers of frames in 2nd video clip of the pair, when the face in a frame is associated with the correct person (i.e. the one seen in the 1st video clip of the pair), without any association with other seen persons. - best (non-hesitant) case

S11 - ... when the face is not associated with one individual, but rather with several individuals, one of which is the correct one. - good (hesitating) case

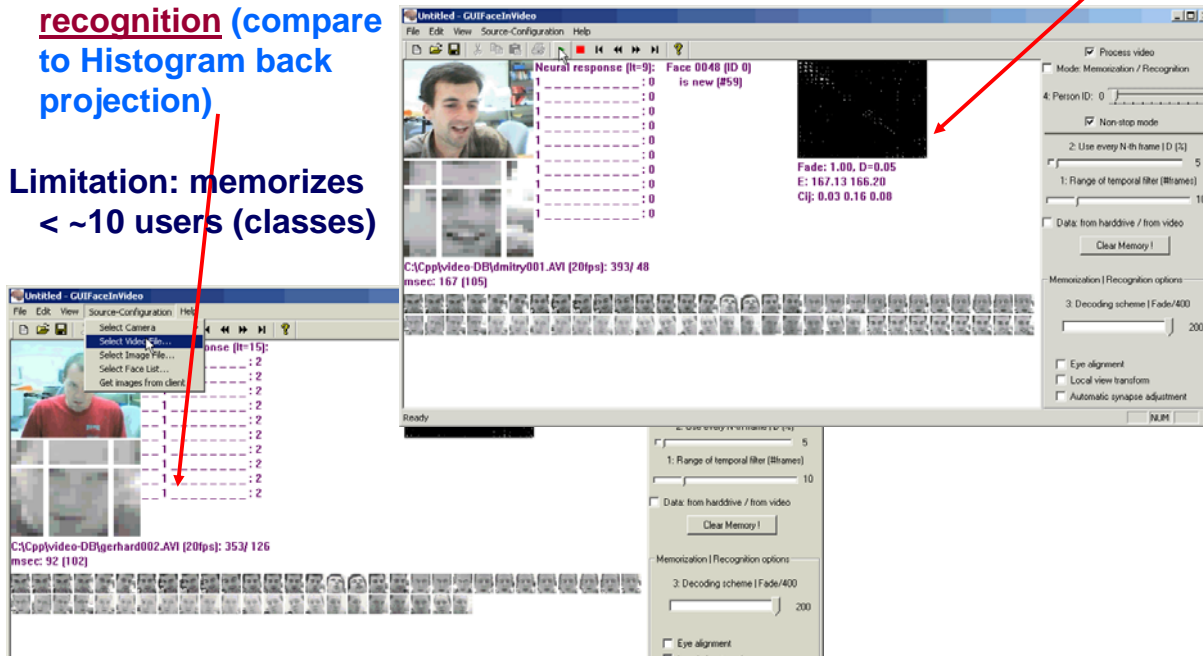
S01 - ... when the face is associated with someone else - worst case
S02 - ... when the face is associated with several individuals (none of which is correct) - wrong but hesitating case

S00 - ... when the face is not associated with any of the seen faces - not bad case

DEMO 1: User / Actor recognition

- Downloadable: Works with your web-cam or .avi file
- Shows tuning of name-stimulus associations (synaptic matrix) as you track a face in memorization (compare to 2D Histogram image)
- Shows associative recall (neurons spikes) as you track a face in recognition (compare to Histogram back projection)

Limitation: memorizes
< ~10 users (classes)



Tested using TV video and database

- TV programs annotation (Recognize 4 party leaders in a show)
- IIT-NRC 160x120 video-based facial database (one video to memorize, another to recognize)

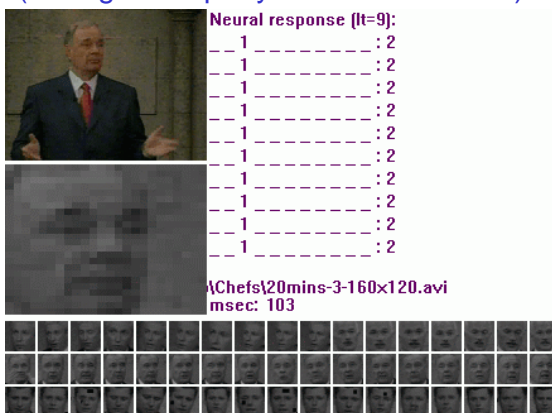
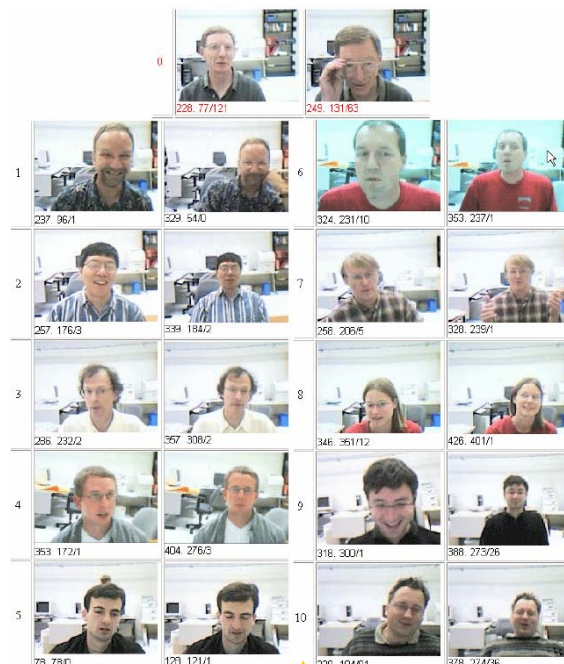


Table 1: Frame-based recognition results.

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



DEMO 2: Squash game

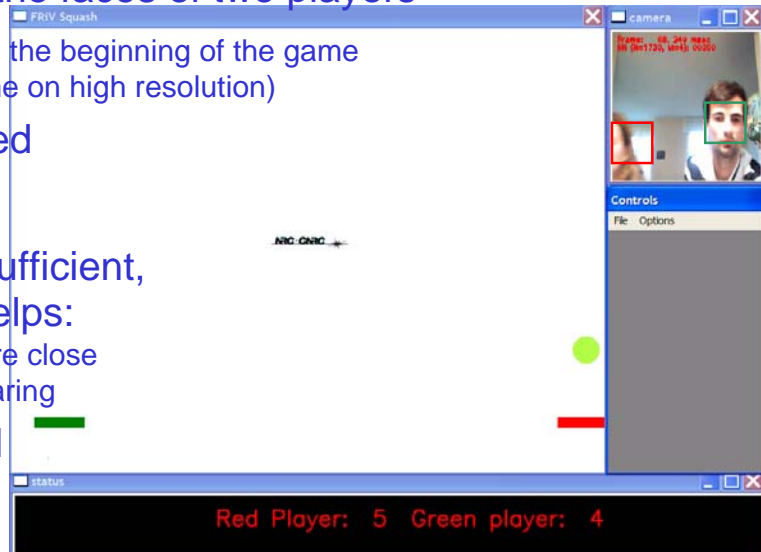
Two-face associative tracker.

(Face motion controls the squash racket (rectangle) to bounce the ball)

Setup: ~1 meter from camera. Game starts when two faces are detected.

- Extension of the face memorizing/recognizing demo to the case of two classes – the faces of two players
- Faces are first learnt in the beginning of the game (slow, due to FD in every frame on high resolution) then individually tracked by the learnt colour histogram reprojection (fast)
- When tracking is not sufficient, associative memory helps:
 - when reprojection areas are close
 - after occlusion or disappearing
- Synapses are updated
 - once in a while
 - when both are detected

15



More details

- Acknowledgements:
 - “Squash” game is implemented by Jason Salares (Carlton University)
- References:
 - Dmitry O. Gorodnichy, **Editorial: Seeing faces in video by computers.** *Image and Video Computing*, Vol. 24, No. 6 (Special Issue on Face Processing in Video Sequences, Editor: D.O. Gorodnichy), pp. 551-556, June 2006.
 - Dmitry O. Gorodnichy. **Video-based framework for face recognition in video.** Second Workshop on Face Processing in Video (FPIV'05) in Proc. of CRV'05, pp. 330-338, Victoria, BC, Canada, 9-11 May, 2005.
 - D.O. Gorodnichy, **Projection learning vs correlation learning: from Pavlov dogs to face recognition.** AI'05 Workshop on Correlation learning, Victoria, BC, Canada, May 8-12, 2005.
- **Open Source Associative Memory Library:**
<http://synapse.vit.iit.nrc.ca/memory/pinn>