NRC·CNRC
From Discovery to Innovation...

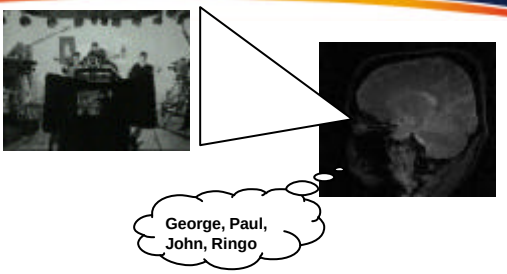
Face recognition in video as a new biometrics modality & the associative memory framework

Talk for IEEE Computation Intelligence Society (Ottawa Chapter)
Dec 15, 2004

Dr. Dmitry Gorodnichy
Computational Video Group
Institute for Information Technology
National Research Council Canada
http://www.iit.nrc.ca
http://synapse.vit.iit.nrc.ca

National Research Council Canada | Conseil national de recherches Canada | Canada

Seeing and Recognizing. Test 1



George, Paul, John, Ringo

How easy it is for us, and difficult for computer!..

Outline

1a. Why in video?
- **What's the difference: video vs photos?**

1b. Vision phenomenon:
- Young area of science for the mixture of expertise
- Why humans are so efficient in face processing in video

1c. Face processing by computers
- *Perceptual Vision* technology
- from detection & tracking, to memorization & recognition

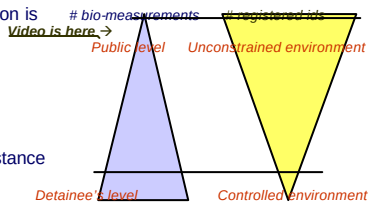
2. Biologically motivated approach to face memorization & recognition in video
- How human brain does it: from retina to synapses
- Demo: NRC-CNRC's "mini models of brain"

Why in video?

Hierarchy of affordability / applicability of different biometrics modalities
(from NATO Biometrics workshop, Ottawa, Oct.2004)

Video-based information is
- most affordable
- least intrusive

- Video provides:
  - soft biometrics
  - identification at distance



Video vs. photographs

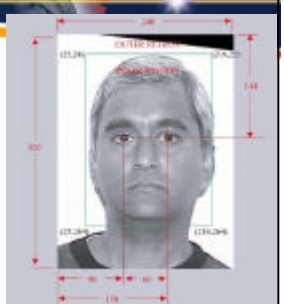
Photos:

Taken in controlled environment

(in procedure similar to fingerprint registration)

- high resolution: 60 pixels **i.o.d.** (intra-ocular distance)
- high quality
- face "nicely" positioned

Stored in databases as such ➔



Canonical face model adopted by ICAO'02 for passport-type documents

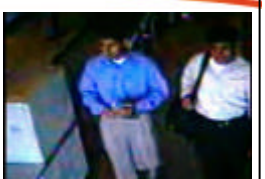
Video vs. photographs (cntd)

Video:

Taken in unconstrained environment.

(in a "hidden" camera - like setup)

- People
  - don't not look into camera
  - even don't face camera
- Poor illumination
- Blurriness, bad focus
- Individual frames of poor quality



Hijackers of 11/9 captured by an Airport surveillance camera

Video can also be MPEG compressed (for storage & transmitting)

Video vs. photographs (cntd)

Face Resolution: 1/16 – 1/32th of the image

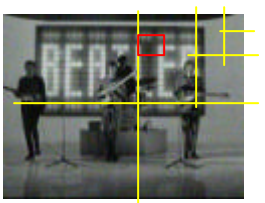
In TV & cams: 640 x 480

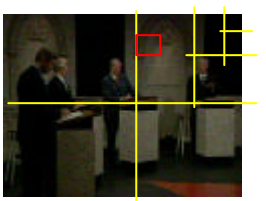
Head = 20-40 pixels

i.o.d. = 10-20 pixels !

- Yet, is this video information really so bad?
- For humans, it is not!

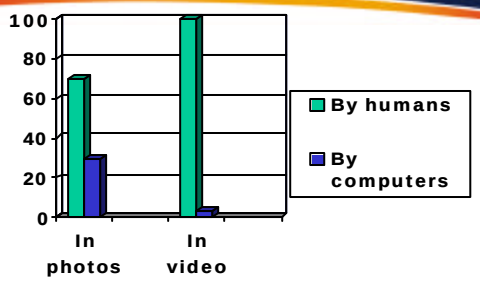
Common face-in-video setups





- Humans don't have problem recognizing faces in video under these "bad" conditions.

Recognizing faces



Refs: [Biometrics forums]

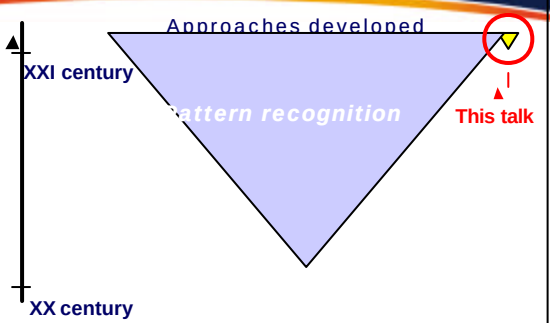
What does that mean?

- For researchers
  - approaches
  - testing benchmarks
- For applications / customers:
  - standards
  - resource allocation

Part 1.
Vision phenomenon.

Face-in-video tasks

Pattern recognition vs Video recognition



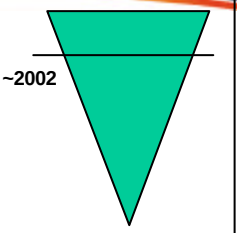
History of video

- Beginning of XX century:
  First motion pictures (movies) appeared...
  - when it became possible to display video frames fast: 12 frames per second
- Beginning of XXI century:
  Computerized real-time video processing...
  - became possible,as it became possible to process video fast: >12 frames per second

Video-processing is a very young area (while pattern recognition is old)

Computers & Video

- Processors (my computer)
  - 1997: 100 MHz
  - 2000: 300MHz
  - 2002: 1GHz
  - 2004: 2.6GHz
- Video camera:
  - 1997: $1000
  - 2000: $200 (+ USB)
  - 2002: $100 (+ USB2, +IEEE 1394)
  - 2004: $30 Can (+ TV, DVD…)



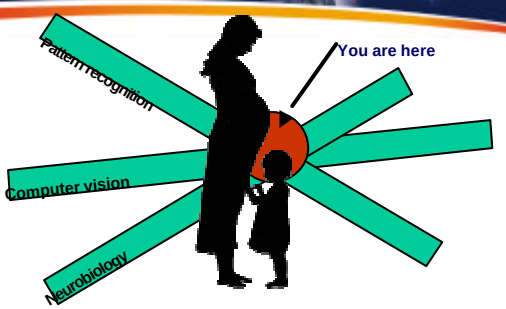
Today: Video has become part of our life.
But only recently it became possible to process it in real-time.

What does that mean?

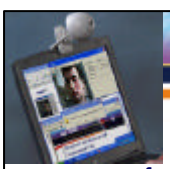
Implications

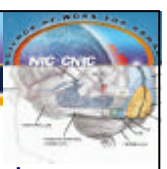
- This area is young and is one of the most rapidly developing
- It still has many unresolved problems: most patents > 2002
- Video-based face processing technology is easily tested – video data is in abundance.
- It involves many stages. Each is important.
- It benefits from the mixture of expertise:

The unique place of this research

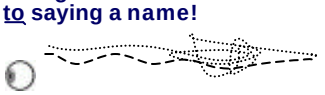


Test 2





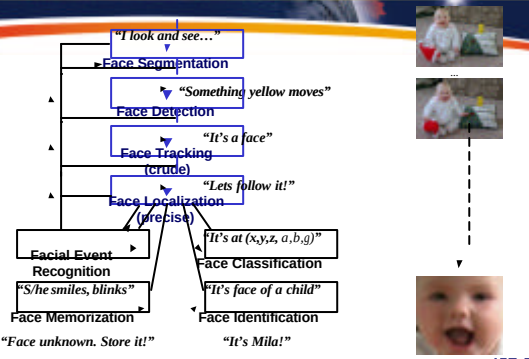
There are many, many stages from seeing a visual stimulus to saying a name!



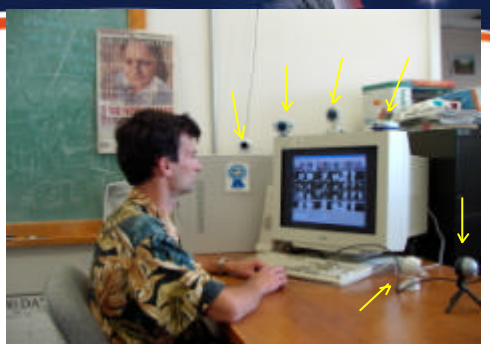
"Jack, Paul, Steven, Gilles"

Can we make the performance of computer systems closer to that of humans?

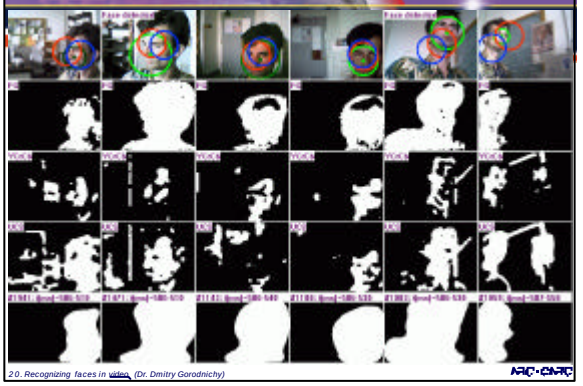
Face processing tasks



Detection and Tracking (demo)



Detection and Tracking (cntd)

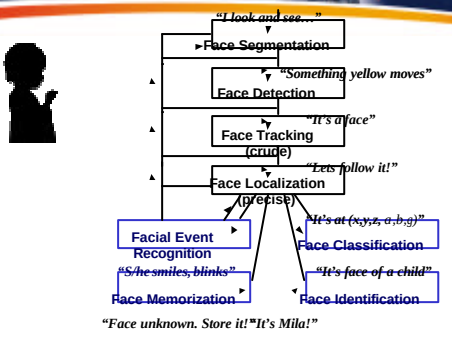


Technology from NRC

For hands-free vision-based computer control

- IP1: Convex-shape feature nose tracking for sub-pixel precision face tracking
  - **Nouse**™ (Use your nose as mouse)
- IP2: Second-order change detection for eye blink detection in moving heads
  - Sending Remote commands to computer
- IP3: With several cameras (Gerhard Roth):
  - 3D tracking with self-calibration NRC kit for cameras

Next stage of perception



Questions to answer

In Memorization of Faces: (assuming it is already detected)

- How to accumulate data over time? Which frames to use?
- Which part of the face to use? What's the resolution?
- What are the features?
- What are recognition techniques?

In Recognition:

- How to use the information from time domain (to compensate for low-res)

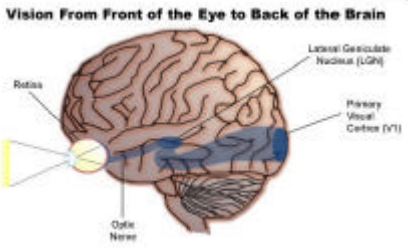
For answers: we'll look into how memorization and recognition of visual information is done in human brain.

For testing: we create a video-based facial database & establish a video-based benchmark

Part 2

Discourse into biological vision systems

How do we see and recognize?



Human brain is a complicated thing …

… but a lot is known and can be used.

The way nature does it

- **Try to recognize this face**

The way nature does it (cntd)

- **What about this one?**

What did you do?

- First you detected face-looking regions.
- Then, if they were too small or badly orientated, you did nothing.
- Otherwise, you turned your face – right?

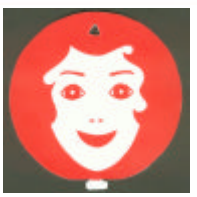
  - This is what biological vision does
  - And what face detectors can do too.

Lesson 1.

- Localization (tracking) of the object first. Then recognition.

  - These tasks are performed by two different parts of visual cortex.

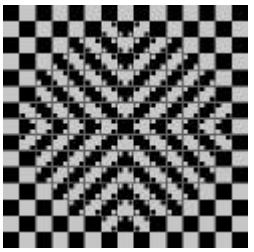
How you beat low resolution



You look only when and where it is needed

Accumulation over time

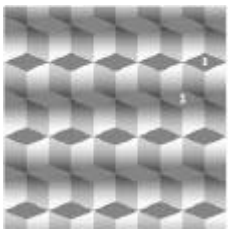
- Fovea vision - Everywhere low quality except we look now.
- Saccadic eye motion
- Accumulation over time



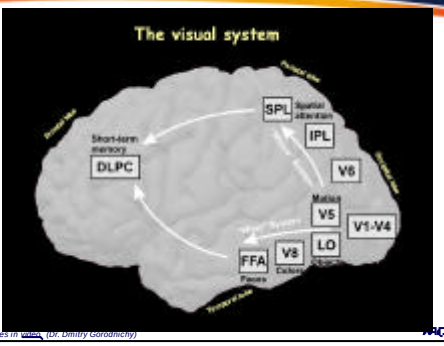
We see what we think we see

- Our recognition decision at time t depends on our recognition decision at time t+1

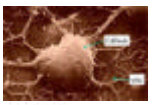
Local brightness adjustment

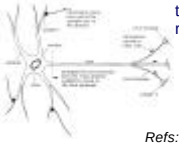


Summary



Biological memory





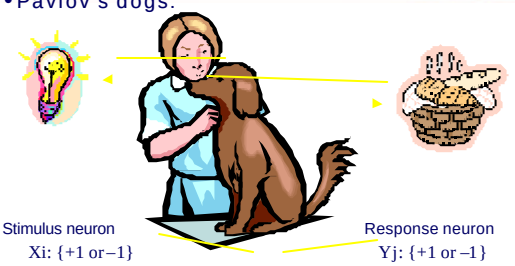
- Brain stores information using synapses connecting the neurons.
- In brain: 10^{10} to 10^{13} interconnected neurons
- Neurons are either in rest or activated, depending on values of other neurons Yj and the strength of synaptic connections: Yi={+1,-1}
- Brain is thus a network of binary neurons evolving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

What we remember are attractors!
This is the associative principle we all live to

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

Main associative principle

- Pavlov's dogs:



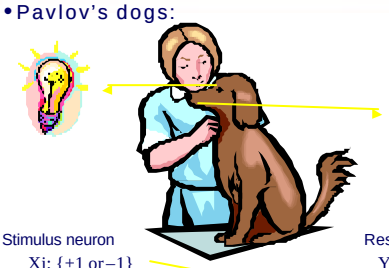
Stimulus neuron
Xi: {+1 or –1}

Response neuron
Yj: {+1 or –1}

Synaptic strength:
-1 < Cij < +1

Main associative principle (cntd)

- Pavlov's dogs:



?

Stimulus neuron
Xi: {+1 or –1}

Response neuron
Yj: {+1 or –1}

Synaptic strength:
-1 < Cij < +1

A Model of memory

- ***Attractor-based neural networks***

$$Y_i(0)=V_{noisy}$$

$$Y_i(t+1)=\mathrm{sign}\Big(\sum_{j=1}^{N} C_{ij}Y_j(t)\Big)$$

$$Y(t+1)=\mathrm{sgn}(CY(t))$$

$$Y(t^*)\ V$$

Main question:

How to compute synapses Cij so that

*a) **the desired patterns V^m become attractors, i.e. $V^m \sim CV^m$***

b) network exhibits best associative (error-correction) properties, i.e.
- largest attraction basins
- largest number of prototypes M stored

?

Refs: Hebb'49, McCalloch-Pitts'43, Amari'71,'77, Hopfield'82,Sejnowski'89…

Facial Visual memory

- Visual memory can be modeled as collection of mutually inter-connected attractor-based neural networks.
- After one network (of features) reaches an attractor, it passes the information further done to the next network (of name tags)

How to update weights

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

From biologically plausible to mathematically justifiable

Neurophysiological Postulate: *"If two neurons on both sides of a synapse are activated, then the strength of the synapse is strengthened"*

"When a child is born, she knows nothing. As she repeatedly observed, she learns" – Postulate from Montessori approach to enfant development.

Models of learning

Hebb: (**C = 1/N VV**^{T}) $\Delta C_{ij}^m = \frac{1}{N} V_i^m V_j^m$, Generalized Hebb: $\Delta C_{ij}^m = aF(V_i^m, V_j^m)$

Better rule: $\Delta C_{ij}^m = aF(C_{ij}^{m-1}, V_i^m, V_j^m)$ or even $\Delta C_{ij}^m = aF(C^{m-1}, V^m)$

Pseudo-inverse as the best rule

$$\mathbf{C = VV^+}$$

- Obtained mathematically from stability condition: $V^m = CV^m$

$$c_{ij}^m = c_{ij}^{m-1} + \frac{(v_i^m - x_i^m)(v_j^m - x_j^m)}{\varepsilon^2} \qquad \varepsilon^2 = |V^m - C^{m-1}V^m|^2$$

- With reduced self-connection ($Cii = 0.15\,Cii$), it is guaranteed [Gorodnichy'97] to retrieve M=0.5N patterns from 8% noise
  M=0.7N patterns from 2% noise (for comparison: Hebb rule stops retrieving when M=0.14N)
- Widrow-Hoff's (delta) rule is the iterative approximation of it.

$$c_{ij}^{m+1} = c_{ij}^m + \alpha(v_i^m - x_i^m)v_j^m$$

- Hebb rule is the special case of it for orthogonal prototypes.

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

It is fast

… besides that it yields the best retrieval for this type of networks.

- It is non-iterative – good for fast (real-time) learning
- It is also fast in retrieval.

[Gorodnichy&Reznik'94]:

"*Process only those neurons which change during the evolution*"
→ instead of N multiplications only a few of them (like in brain)
→ very few iterations in retrieval

It converges to attractor

- **Guaranteed to converge.**

The behaviour of the network is governed by the energy functions

$$E^{Hopf}(Y(t)) \doteq -\frac{1}{2}\bar{Y}^t(t)C\bar{Y}(t) \qquad E^{dyn}(Y(t)) = -\frac{1}{2}\bar{Y}^t(t+1)C\bar{Y}(t)$$

The network always converges: as long as Cij=Cji [Gorodnichy&Reznik'97]

It deals with flow of data

- It can deal with continuous stream of data, never being saturated, just like real brain does.

  → maintaining the capacity of 0.2N (with complete retrieval)

  → providing means for dealing with flow of data

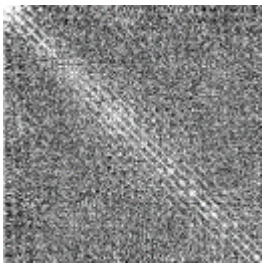
How do stored data look like?

The look inside the brain:

By looking at the synaptic weights *Cij*, one can say what is stored:

- how many main attractors (stored memories) it has.

- how good the retrieval is.



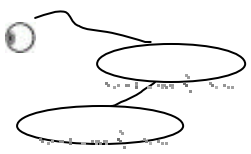
Summary

... All this makes the pseudo-inverse neural network very suitable for continuous real-time video processing.

- This is the network used to build "mini models of brain" for memorizing and recognizing faces from video sequences.

-

Facial Visual memory



- Visual memory can be modeled as collection of mutually inter-connected attractor-based neural networks.
- After one network (of features) reaches an attractor, it passes the information further done to the next network (of name tags)
- Neurons responsible for specific tags will fire

Putting it all together

Using the described ideas we now build visual memory...

...Under these constraints:

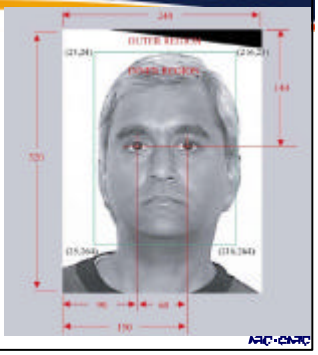
- Real-time processing is required.
- Low resolution: 160x120 images or mpeg-decoded.
- Low-quality: week exposure, blurriness, cheap lenses

Video-based facial database is needed for performance evaluation – *How to build it?*

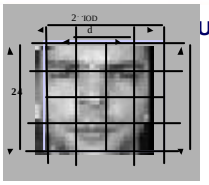
Video-based DB: Canonical model

Model adopted for Face Recognition in passport-type documents [ICAO'02]

- Is it suitable for Face Recognition from video? – no
- Is this the most informative face representation (for humans)? – No.
  - No 3D information [fpiv04].



Video-based DB: features



Optimal set of features-

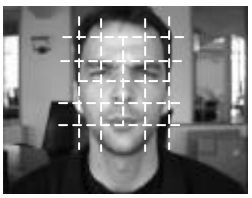
Use lowest resolution possible, not to inflict overfitting due to the noise.
And there's a lot of noise in video!

Size 24 x 24 is sufficient
(for face memorization & recognition and is optimal for low-quality video and for fast processing)

And you can use many of those:
a video sequence instead of one image.

Applicability of 160x120 video

According to face anthropometrics
-studied on BioID database
-tested with Perceptual Vision interfaces
-observed in cinematography



Face size	½ image	¼ image	⅛ image	1/16 image
In pixels	80x80	40x40	20x20	10x10
Between eyes-IOD	48	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	√	√	-	-

√– good
b – barely applicable
- – not good

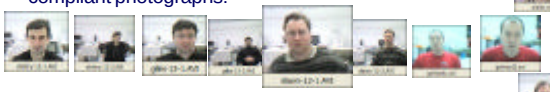
What can be detected

The range of facial orientation and expression tolerated by current face detection technology
- can find faces as long as **i.o.d >10 pixels**
- in bad illumination,
- with different orientations,
- facial expressions

- If face is too far or badly oriented:
  - Motion and colour can be used to detect and track.

Video-based performance evaluation

For recognition in video:
- These 160x120 (20 secs) mpeg-ed video files (1.5Mb) showing facial motion are more informative than ICAO-compliant photographs!



NRC's database consists of two video clips of 10 member of CV group (one used for training, the other for testing)

- They can be used for testing and as a benchmark.
- However, any video can be used as a benchmark
  – From TV show, movie, internet

General conclusions

- FR in Video is a new biometrics modality.
  - **It is NOT part of (and not to be confused) with FR in Photos**
  - **FR in Photos (or in controlled video environment) is "hard" biometrics like Fingerprints**
  - **FR in Video (in uncontrolled video environment) is "soft" biometrics like height.**
  - **FR in Video is most powerful of all soft biometrics**
- **It means:**
  - **New (different) standards, benchmarks, approaches**
  - **some shown in this talk**
- FR in Video is new area & has a lot of potential
  Proof: "If you can recognize a person, computers should be able to do it too" – **This is the ultimate inspiration and benchmark.**

Specific conclusions

- **Inner square part of the face is most important**
- **12 i.o.d. , which is most frequent case, is sufficient for recognition!**
  - **24x24 face area can be extracted in each frame**
- **Archive faces as video clips (not in photographs)**
- **Test approaches on video clips of low-resolution**
- **Human brain (attractor neural network) based approach offers new direction for finding solution.**
  - **Is ready to use in small-environment scenarios (Talk-shows)**
- **Also, vision allows hands-free control/interaction,**

 Now some demos!

Demos

- Demo with a pre-recorded talk show:
  - Memorizing and Recognizing our Political Party leaders in a TV show.
  - 4 persons
  - 15 secs video clips showing each leader are used for memorization.
  - 30 mins video of the debates is used for testing the quality of recognition
- Demo with NRC-CVG facial database
- Live Demo with the audience.
  - Memorizing new faces right now
  - Recognizing them (and still recognizing "Leaders")

