Face recognition: from contactless shopping to advanced forensics, challenges and opportunities

Ebroul Izquierdo

Multimedia and Vision Group

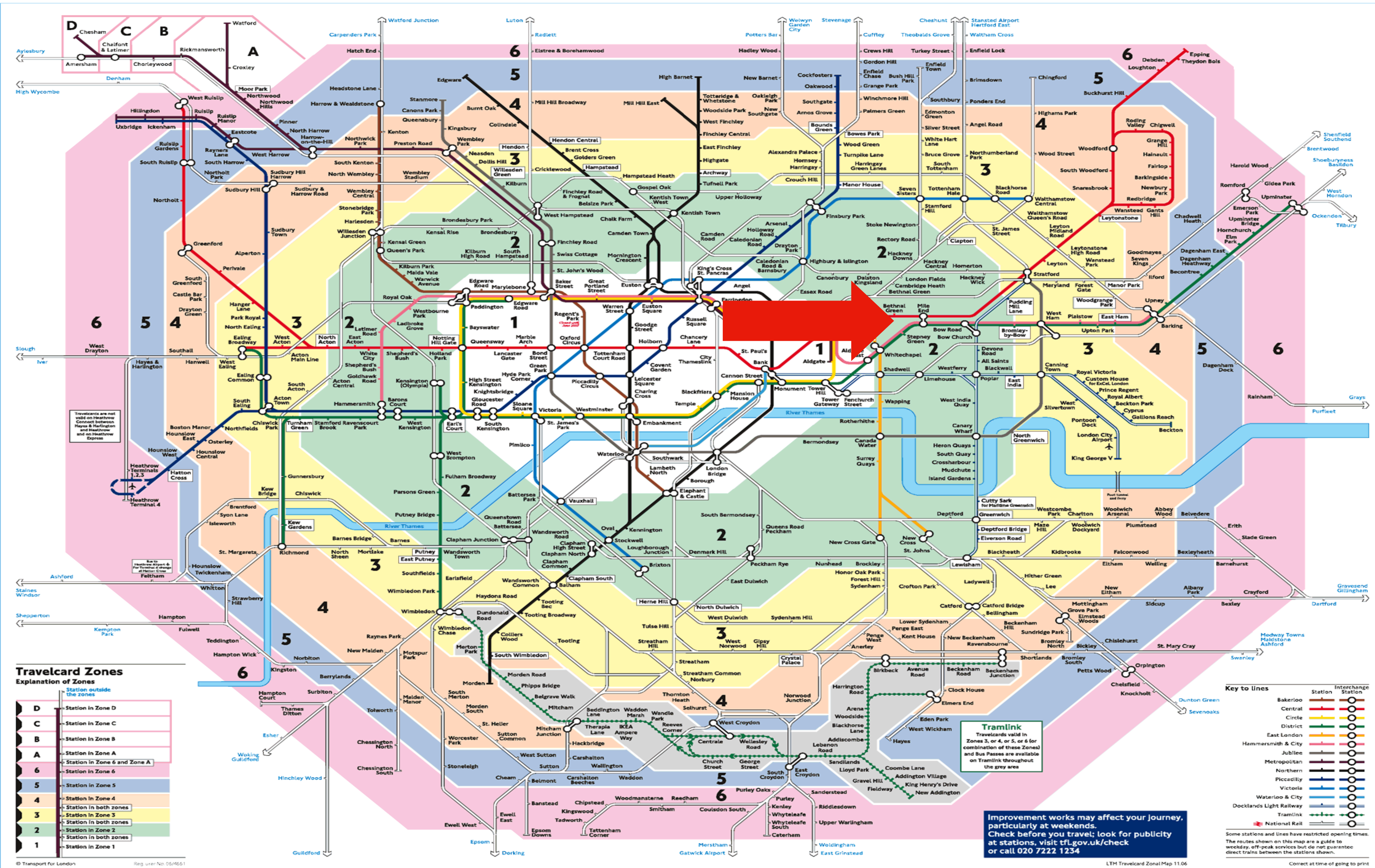
Queen Mary, University of London

JIAP, Montevideo, Sep. 1th 2015

Queen Mary University of London

Multimedia and Vision Research Group


MMV
Multimedia & Vision Research Group



MAYOR OF LONDON

24 hour travel information 020 7222 1234

Website tfl.gov.uk

Textphone 020 7918 3015

Transport for London

QMUL

QMUL & EECS

- Second largest college of the Federal University of London (after University College)
- 4 Campuses across London with Mile End hosting the EECS
- Unique location in the east end (full of character)

EECS

- 60-70 academic staff including 14 professors/Chairs
- 550 undergraduates, 210 MSc, 130 PhD students

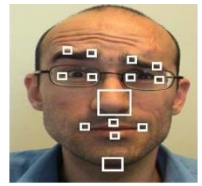
Prof. E. Izquierdo
Dr. Y. Patras
Dr. H. Gunes
Dr. P. Hao
Dr. Q. Zhang
8 Postdoc Ras
28 PhD students

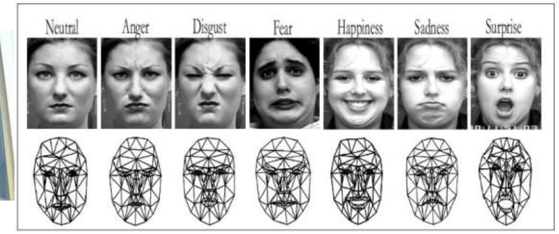
Queen Mary University of London
MMV
Multimedia and Vision Research Group

Looking at people

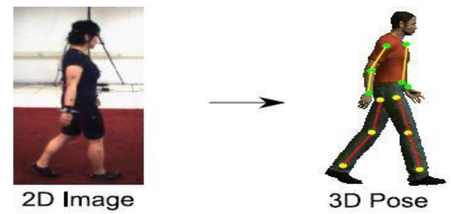
Applications: Multimedia content analysis,
Man Machine Interaction, Smart spaces (e.g. assisted living, health care)

- **Expression Analysis**
  - Head and feature tracking
  - Facial expression recognition



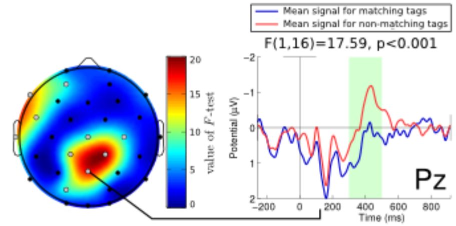


- **Action and Gesture Recognition**
  - Action recognition and localisation
  - Pose estimation
  - Gait recognition



Human Sensing (EEG, gaze, face/body)

- Multisensor analysis of behaviour
  - Affective/cognitive state recognition



- Learning with humans in the loop
  - Image categorization
  - Implicit tagging

Queen Mary University of London
Multimedia and Vision Research Group

PLUS Factorisation and Applications

The Mathematics

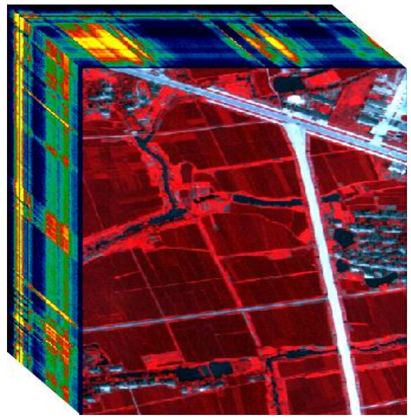
A=PLUS

P : permutation

L : unit lower

U : customisable

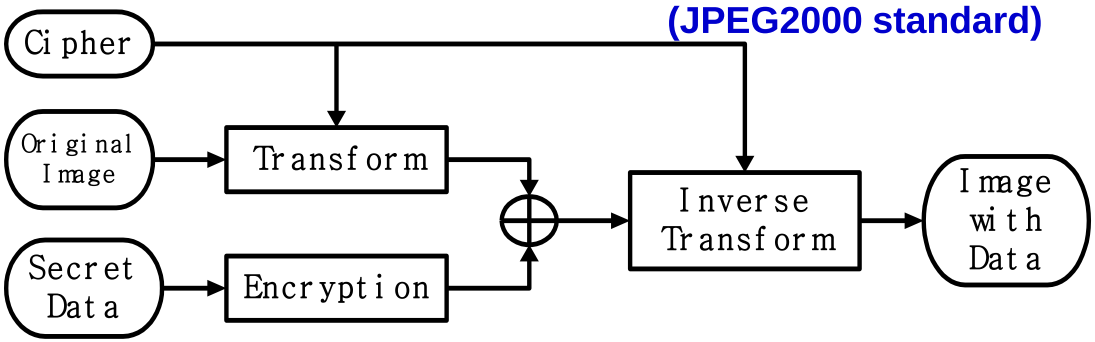
S : special matrix



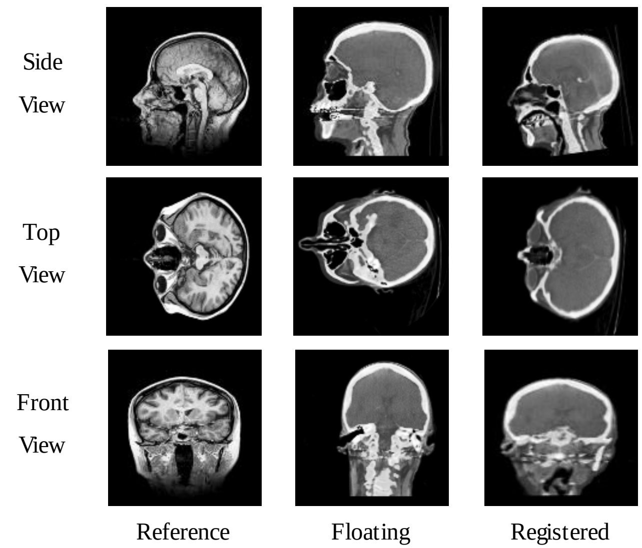
Multiple component image coding : RKLT is the best for both lossy and lossless (JPEG2000 standard)

The Applications

- Lossless image coding
- Fast image transformation
- Huge data hiding
- ……………



Huge data hiding : 70KB in a 512x512 image for 36.37dB loss

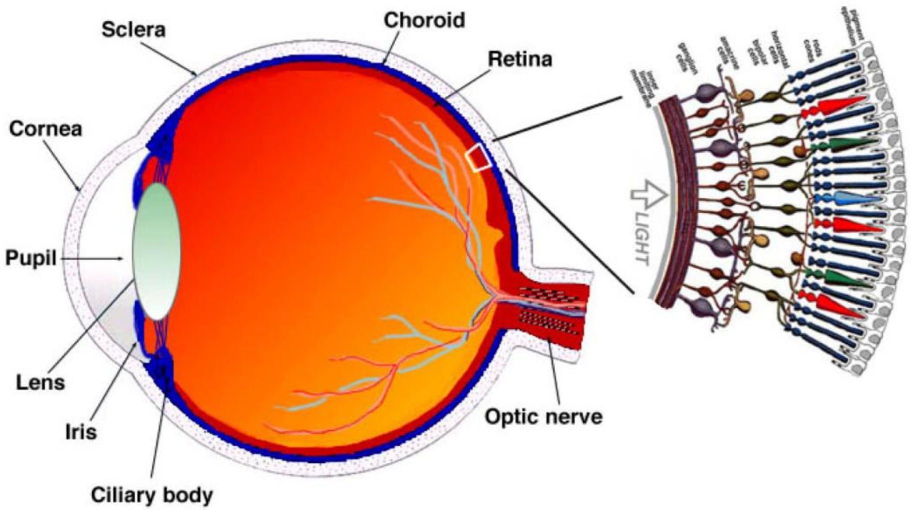


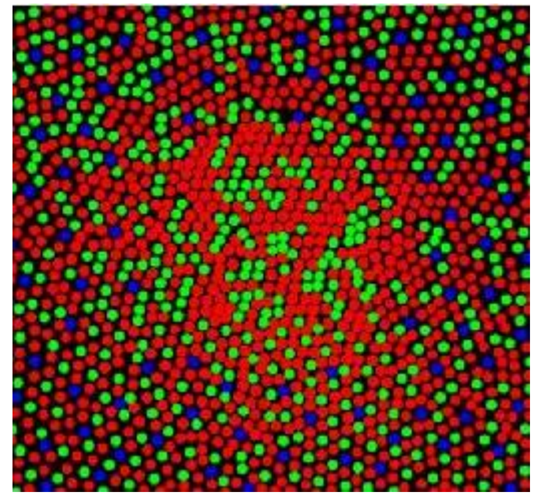
Fast image transform & registration : geometric transforms can be 10 times faster than naïve transformation

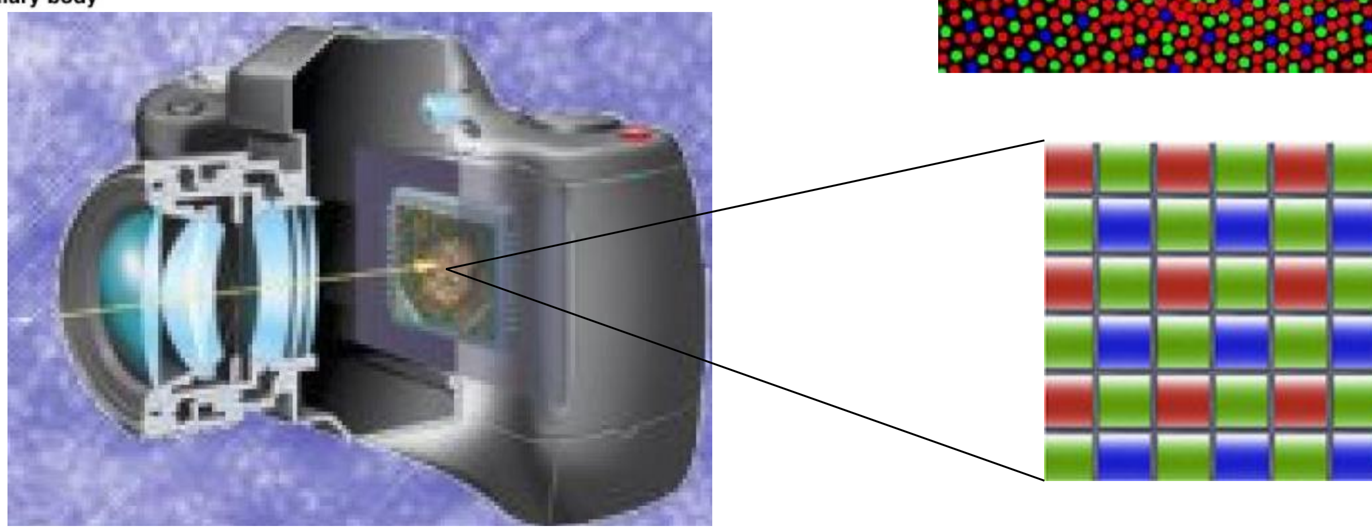
Queen Mary University of London
Multimedia and Vision Research Group

Digital Camera vs Human Vision

Mosaics and demosaicking of colour filter arrays

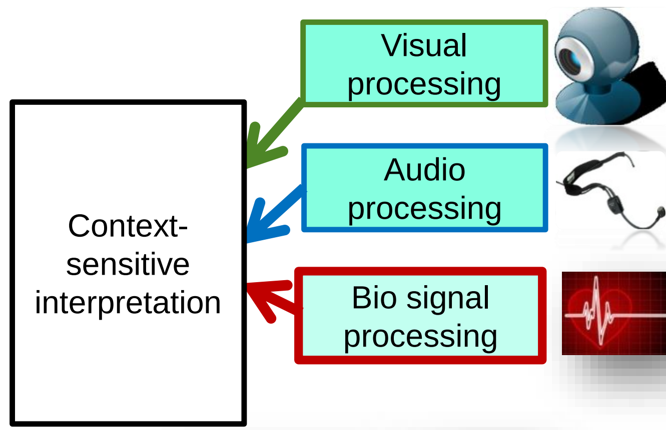


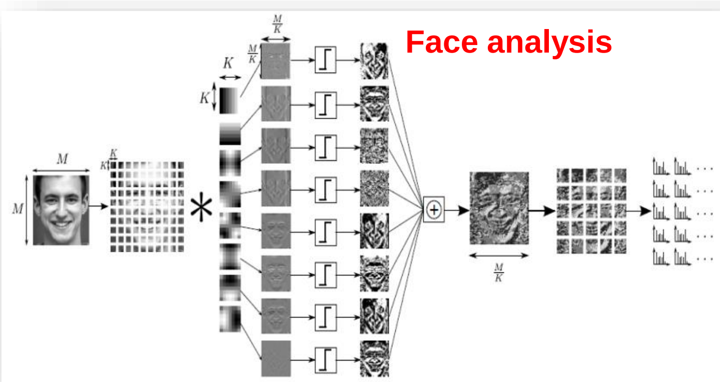




Affective & social signal analysis and understanding

- Affective computing & human behaviour understanding
  - Automatic analysis of
    - nonverbal behaviour (e.g., nodding, smiling)
    - emotions (e.g., arousal, valence)
    - social signals (e.g., personality)
  - Behavioural data acquisition, annotation and representation
  - Affective and social behaviour prediction
  - Multicue and multimodal recognition (fusion)
  - Human-computer interaction
  - Human-virtual agent interaction
- Visual & multimodal information processing
- Machine learning





Biologically Inspired Methods for Image Classification

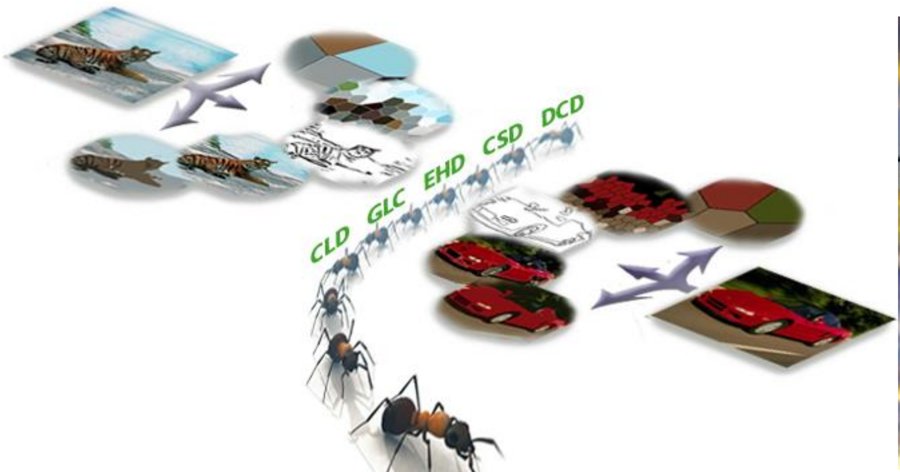
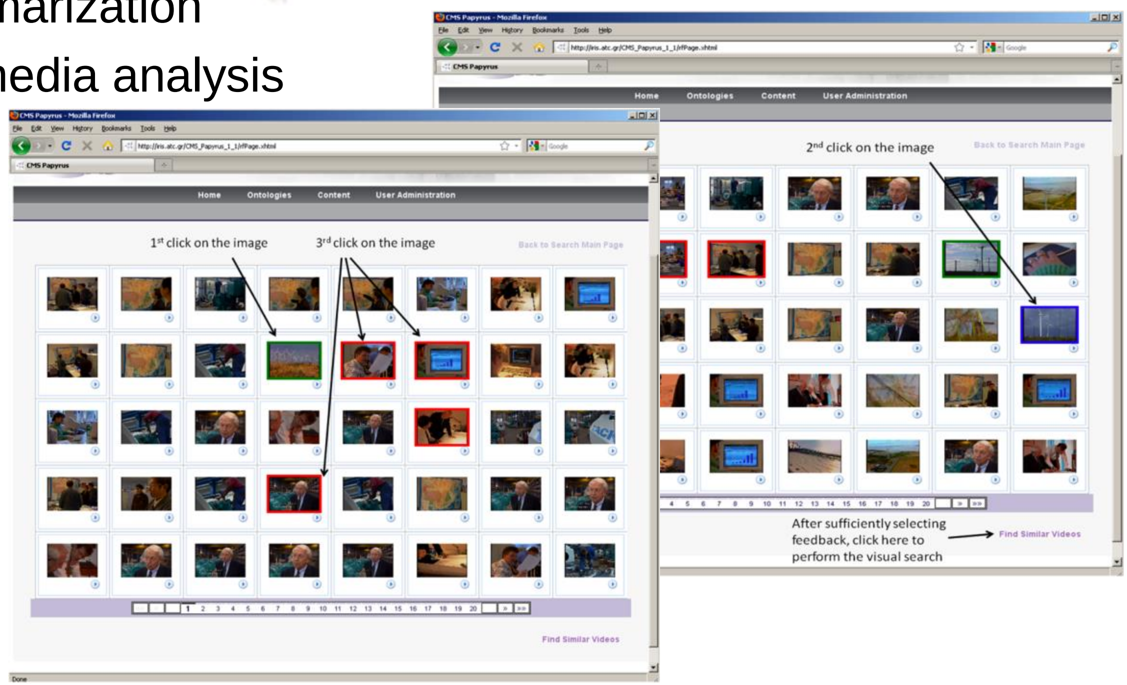


Image and Video classification

Video summarization

Semantic media analysis

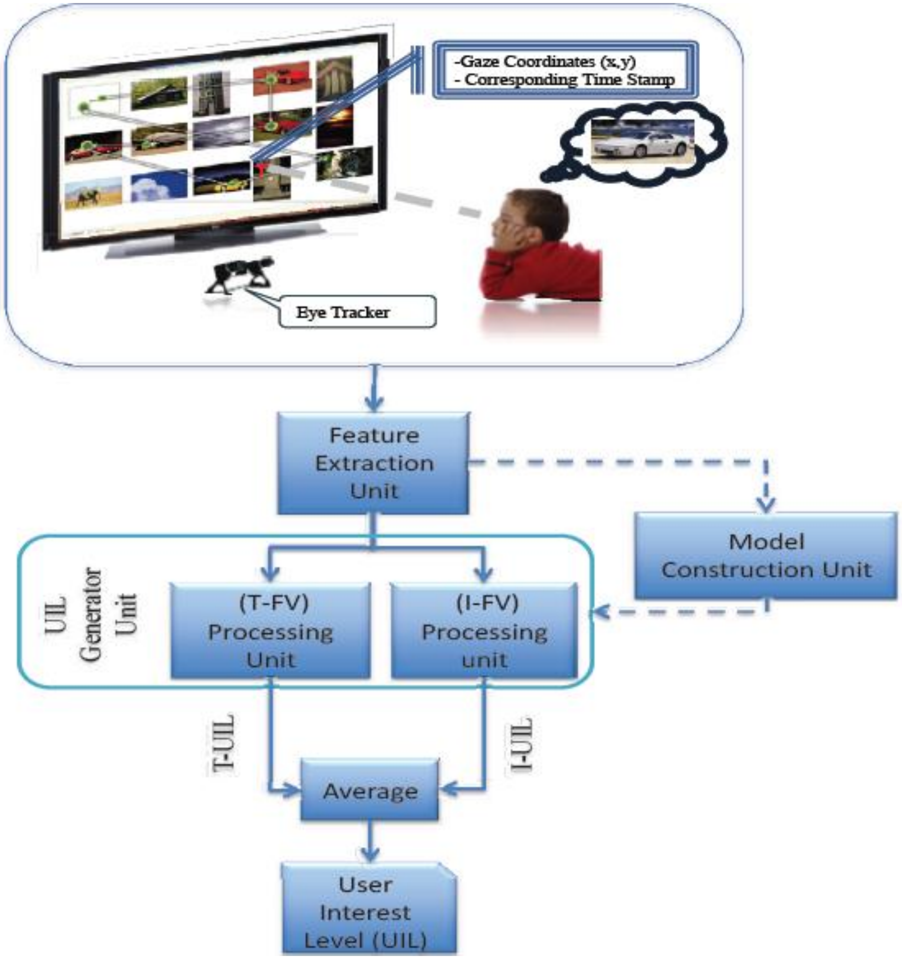


Relevance feedback

Queen Mary University of London

Multimedia and Vision Research Group

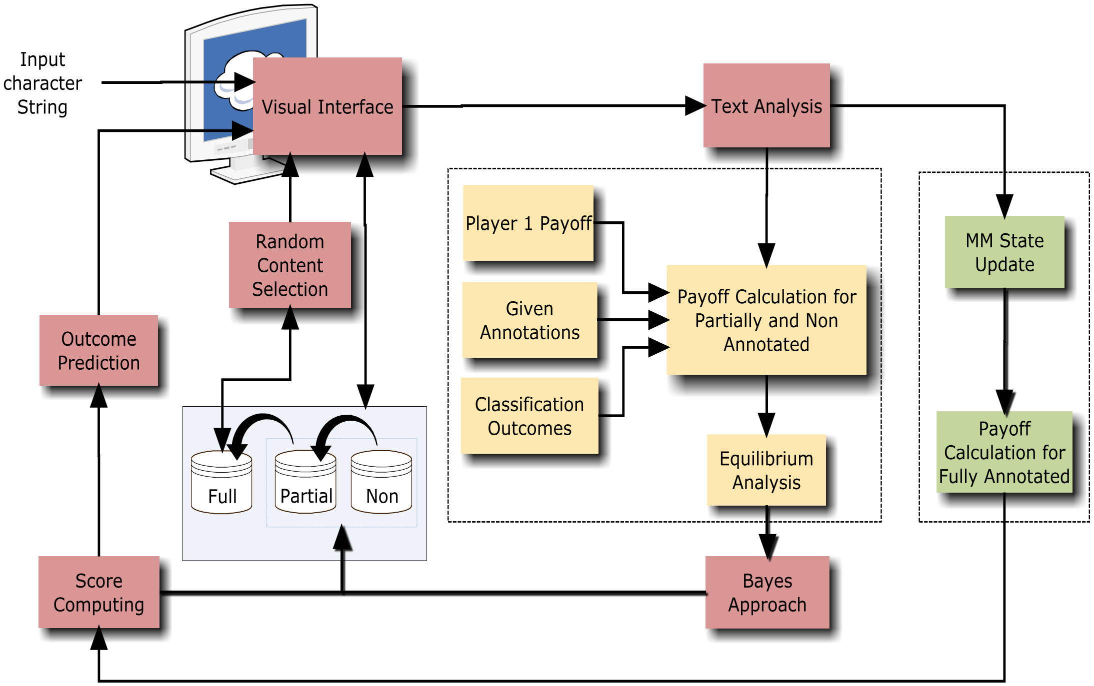
Eye –Tracking

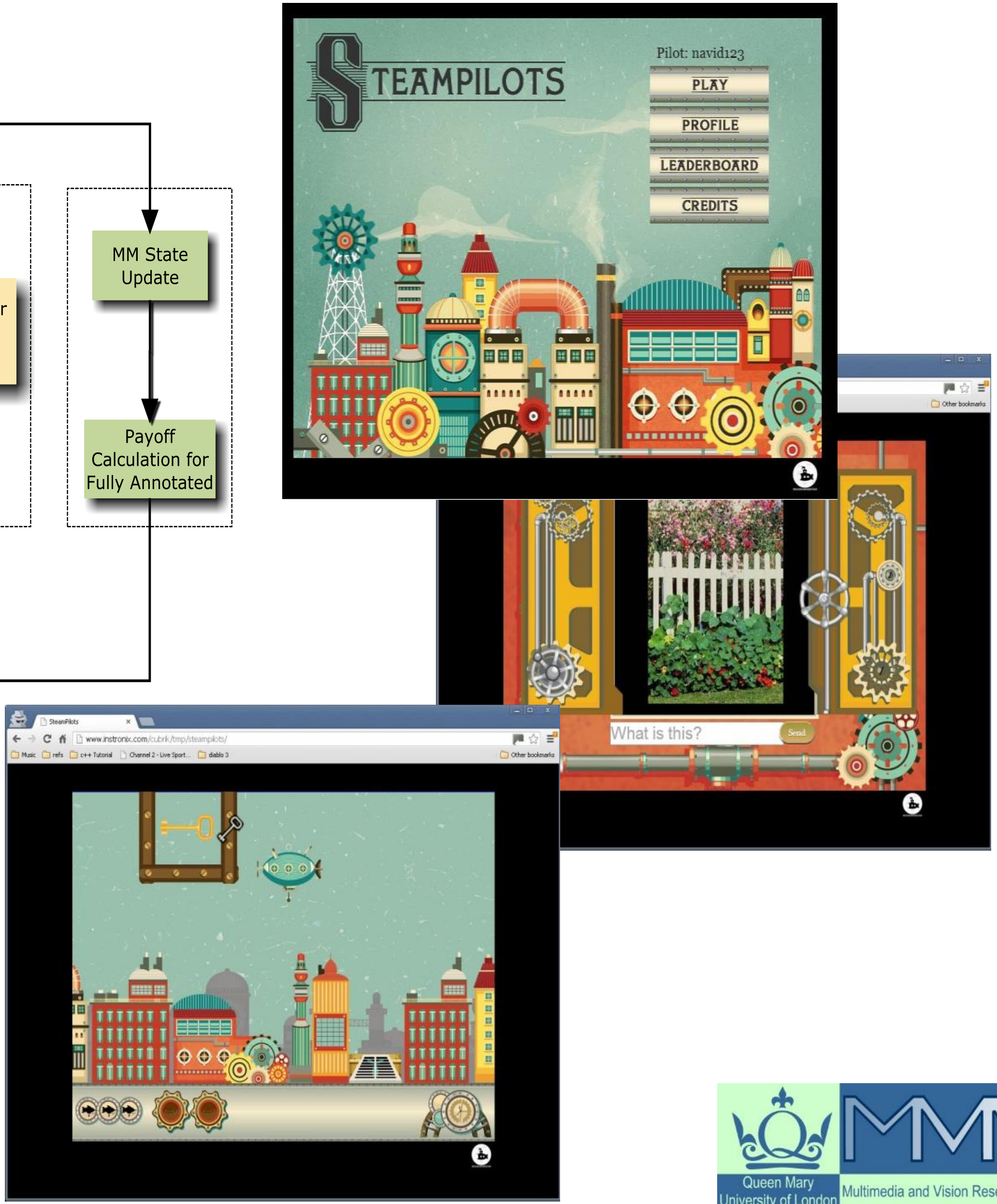


Implicit image tagging

Implicit relevance feedback for search optimization

Serious Gaming





Queen Mary University of London

Multimedia and Vision Research Group

Image-base CAPTCHAs

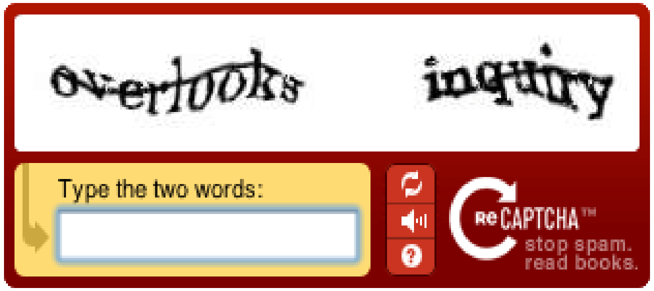
Security in web applications

Increase gap between human & machine recognition

Advanced software for automatic separation

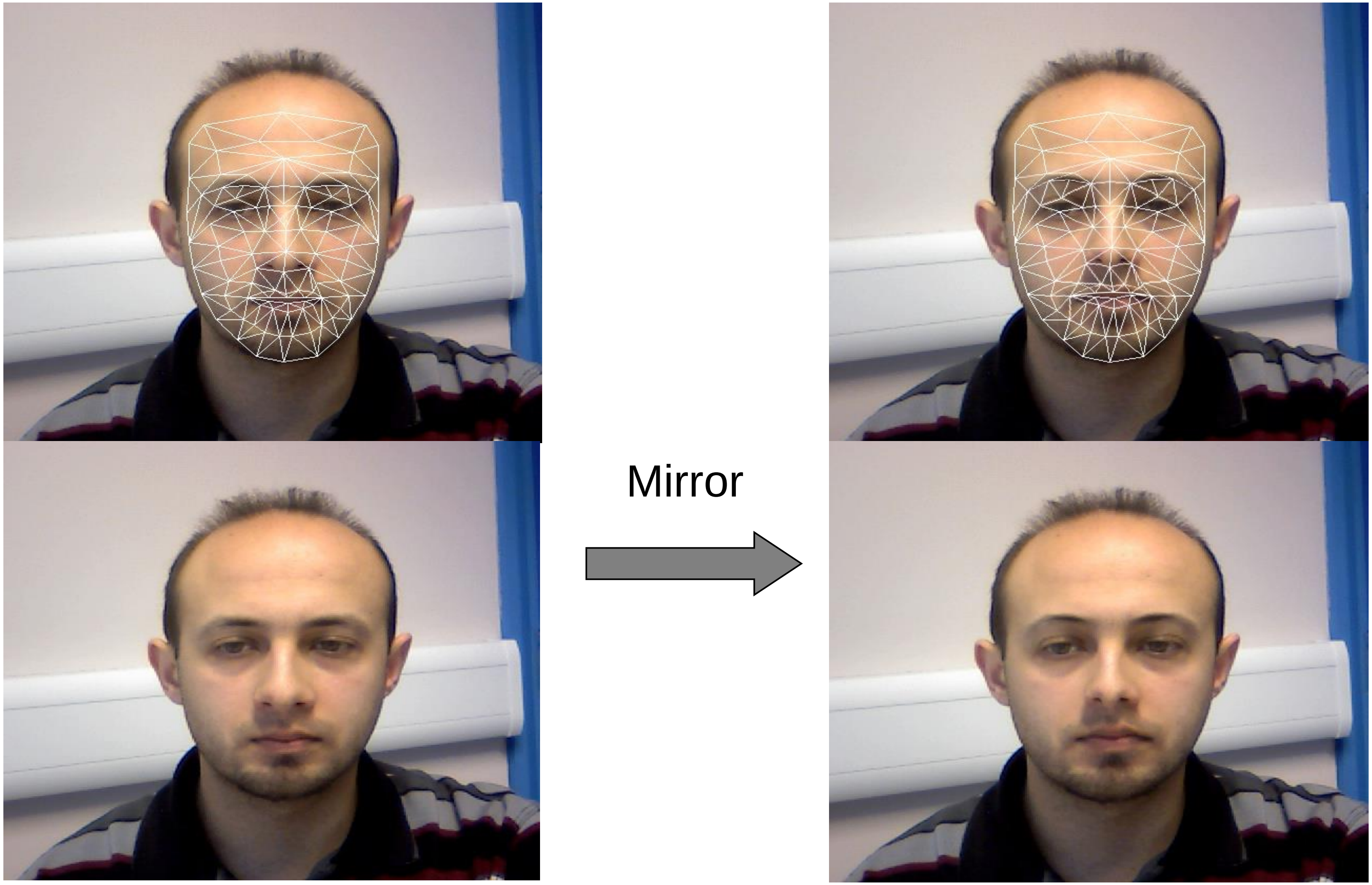
CAPTCHA Test

3D Models



Face Recognition

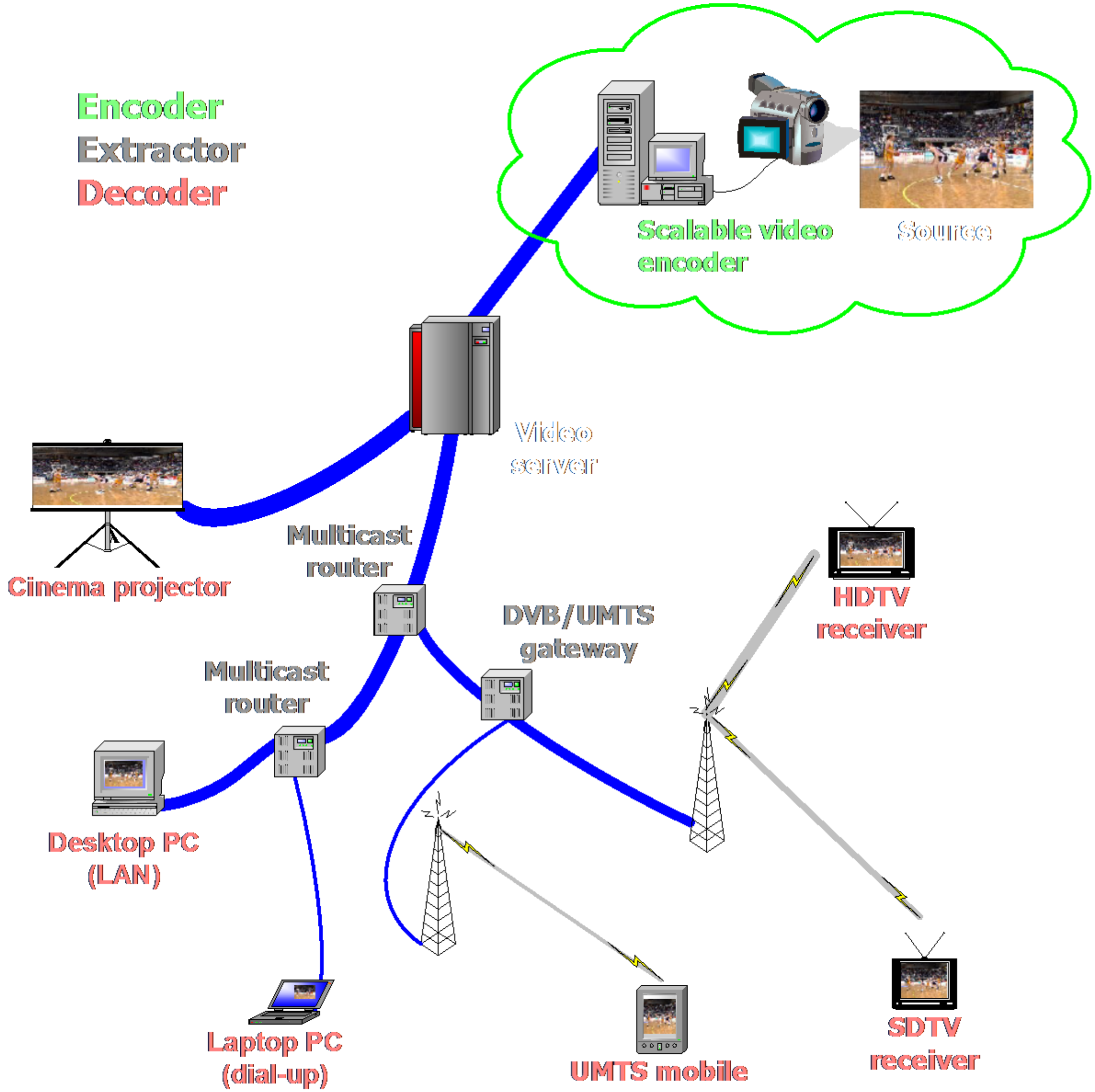
Visualising Changes of Facial Features in Augmented Reality



Queen Mary University of London
Multimedia and Vision Research Group

Scalable Source and Channel Coding Framework

- Server stores the highest quality version of the scalable coded video
- Distribution systems:
  - low bandwidth dial-up
  - ...
  - HDTV broadcast
- Display devices:
  - hand-held
  - ...
  - cinema-quality projectors
- Content can be "upgraded" to a higher quality



Queen Mary University of London
Multimedia and Vision Research Group

Surveillance Centric Coding

Original Sequence

Background Subtraction

Combined adaptation w.r.t event

Temporal adaptation w.r.t event

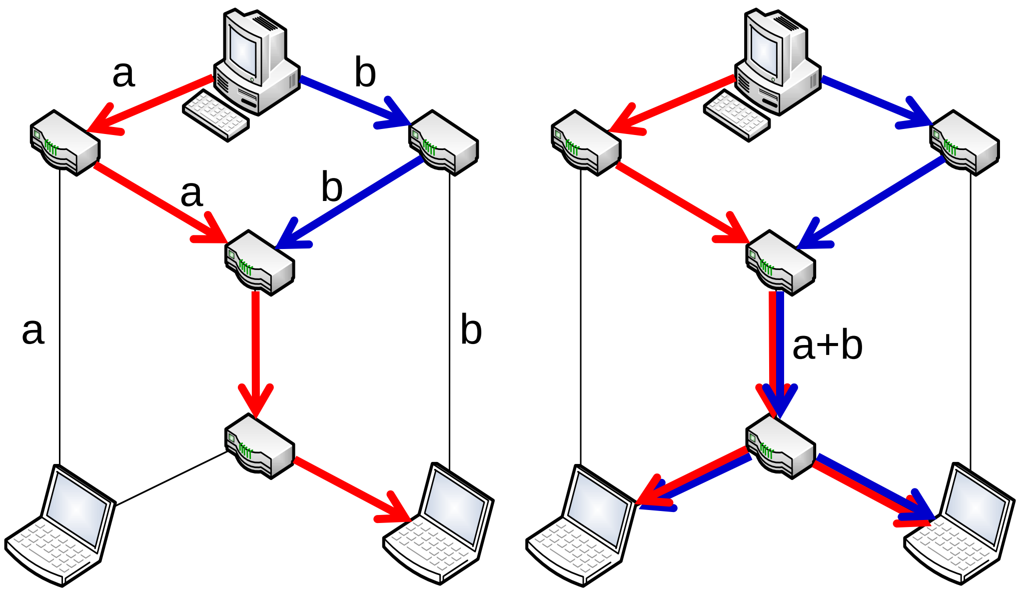
Event detected encoding at high resolution otherwise low resolution

Event detected encoding at high quality otherwise low quality

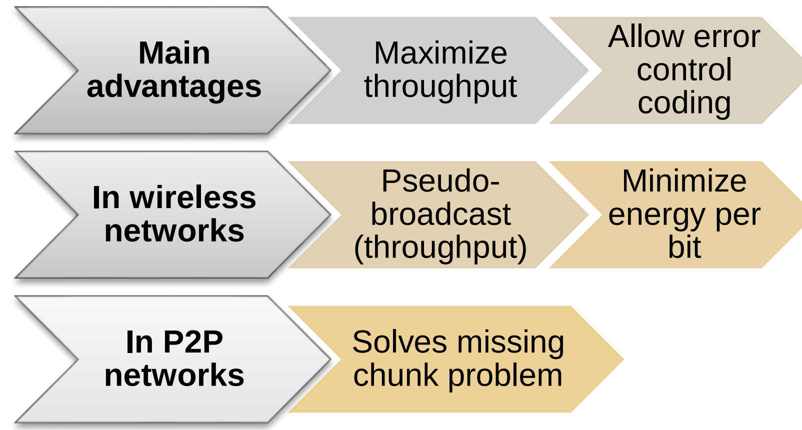
Queen Mary University of London

MMV Multimedia and Vision Research Group

Scalable video streaming with network coding



a) Traditional routing b) Network coding



- Objective: Received video depends on users' bandwidths
  Minimize decoding delay

  Tools: Streaming overlay with P2P push

Queen Mary University of London
Multimedia and Vision Research Group

3D Video Analysis and Mixed Reality

Capture the actual scene

Create the 3D Models

Create the Environment

Share the Experience

Invite your Social Friends

Queen Mary University of London

MMV Multimedia and Vision Research Group

Research Input & Output

- Over the past few years, grants worth
  - £1.2 Million UK Funding bodies
  - £5 Million EU & other sources in 2010-2015
- Output: papers
  - Over 100 peer-reviewed Journals
  - Over 300 peer-reviewed conference papers



FreeFoto.com


© Derek Langley/darknessandlight.co.uk

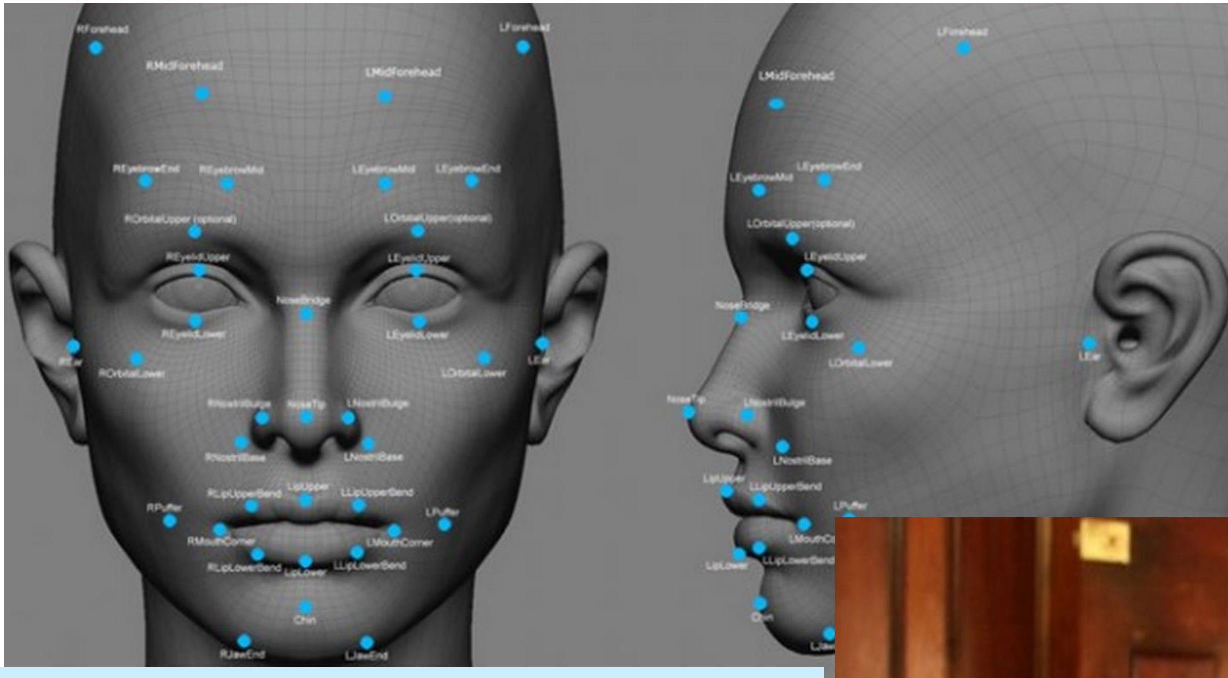

Hackney Central D6
First
DMC42518
Hackney Central D6
DMC42518
SALESALE

The Challenge

How accurate is facial recognition software now?

Queen Mary University of London
MMV Multimedia and Vision Research Group

How accurate is facial recognition software?



“Facebook’s facial recognition, DeepFace, is now very nearly as accurate as the human brain.”

“DeepFace can look at two photos, and irrespective of lighting or angle, can say with 97.25% accuracy whether the photos contain the same face. Humans can perform the same task with 97.53% accuracy.”

How accurate is facial recognition software?

A recent report from The National Institute of Standards and Technology (NIST) state: …"*results from its 2013 test of facial recognition algorithms show that accuracy has improved up to 30 percent since 2010.*

The error rate decreased from 8.9 percent in 2010 to 6.4 percent in 2013."

"It includes results from algorithms submitted by 16 organizations. Researchers defined performance by recognition accuracy—how many times the software correctly identified the photo in a massive photo data sets. "

Queen Mary University of London

Multimedia and Vision Research Group












The Challenge

- **In a surveillance (CCTV) application, how accurate is face recognition?**
- **Given 10000 faces automatically detected or manually cropped from real-world (CCTV) footage, how many faces would state of the art software recognize?**

My initial answer: **8000**

True experimental answer: **18**

Another Challenge

Given a *query face* manually cropped from real-world (CCTV) footage and a database of 20 annotated images containing one or more different pictures of the *query face*,

What is the probability of finding a match?

My initial answer: 0.8

True experimental answer: **0**

The "Query Face"

Obviously the machine fails!

Can the human brain and AV system do a better job?

Face Recognition: Good, Bad and Ugly

- Many interesting applications
  - Search engines: Find pictures of Madonna
  - Surveillance and tracking
  - **Forensics, authentication and security**
- It is a very special case of object recognition
- One of the key applications of VIR
- It is an “easier task” (compared with generic object recognition)
- It relates to the most important known object

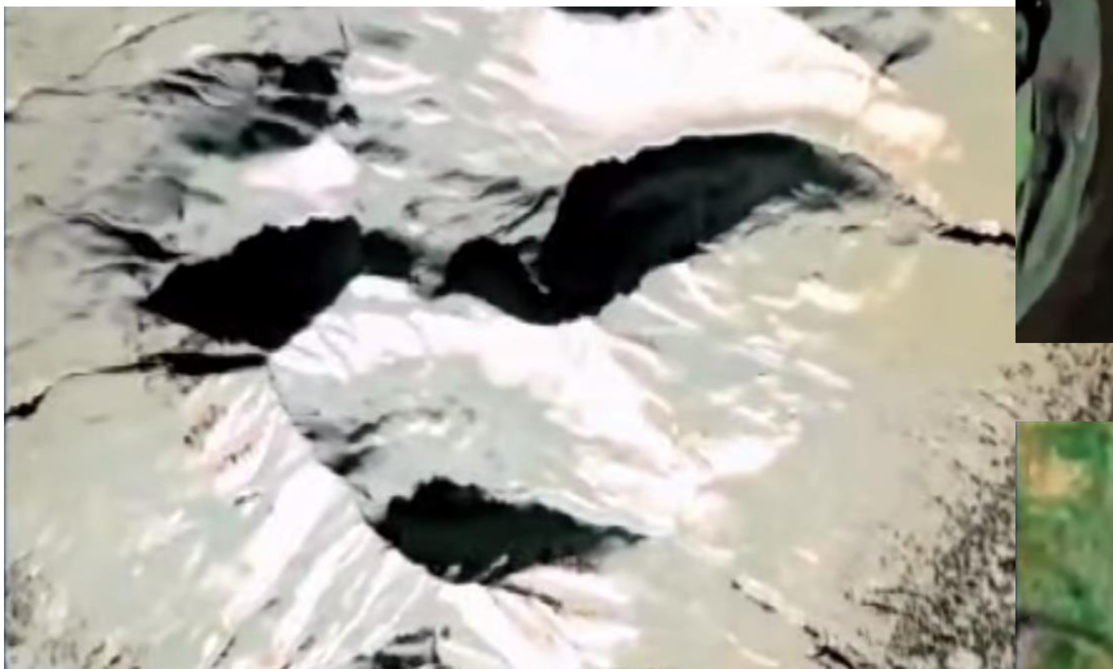
The good

Apparently the human brain/AVS has been "hard-wired" (by evolution) to recognize faces

We can see or recognize faces almost everywhere

The good

Google Faces



Scanning of pictures from Google Earth, looking for “faces”

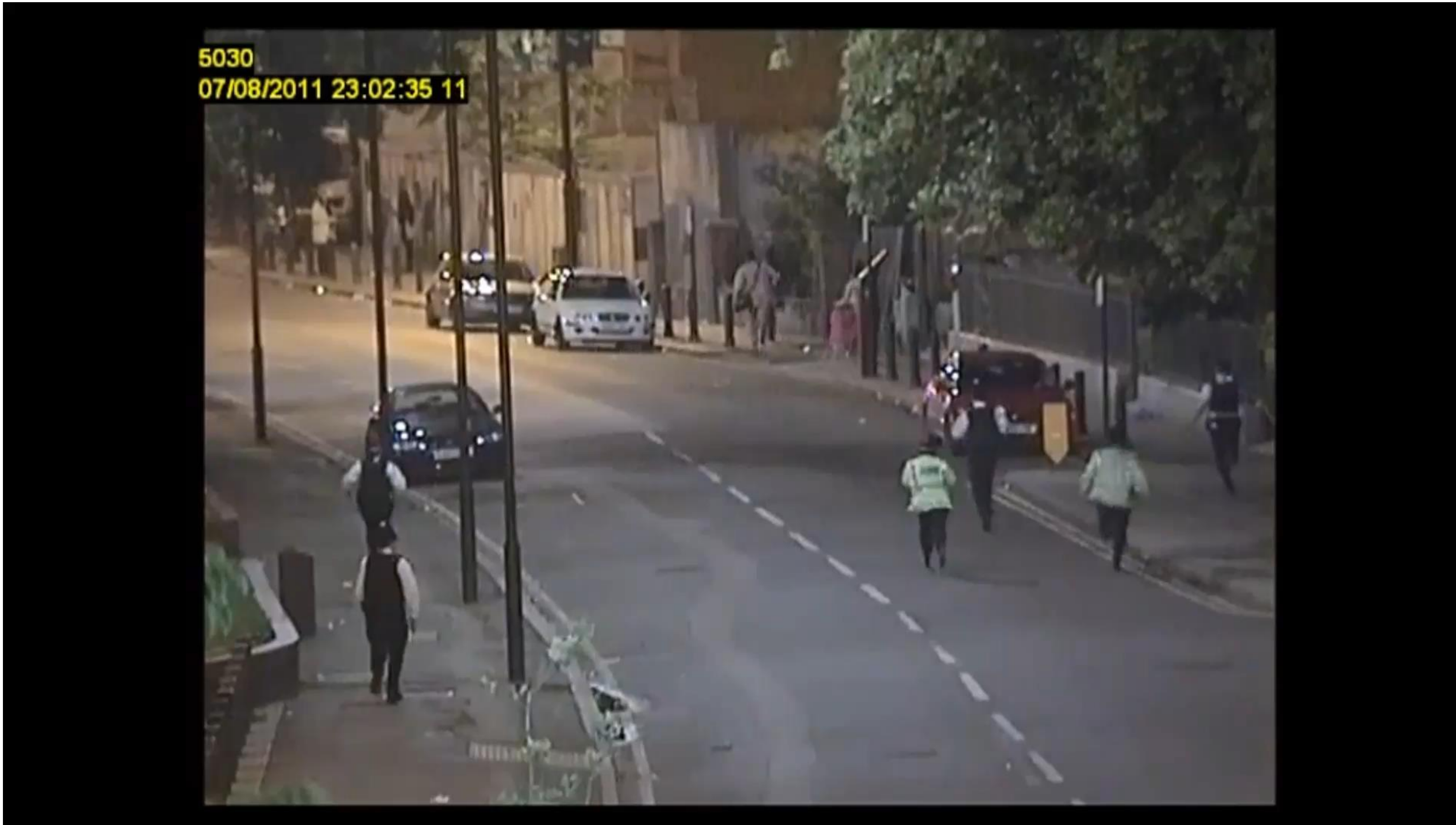
Authentication: No more password and PIN; Cannot be stolen; Cannot leave it at home; ***No more loyalty reward cards***

Queen Mary University of London
MMV Multimedia and Vision Research Group

The bad

- Important commercial applications requires extremely high accuracy
- Seems to breakdown in a critical “real-world” applications related to security/surveillance under adverse conditions

The bad



The ugly: severe privacy and ethical implications

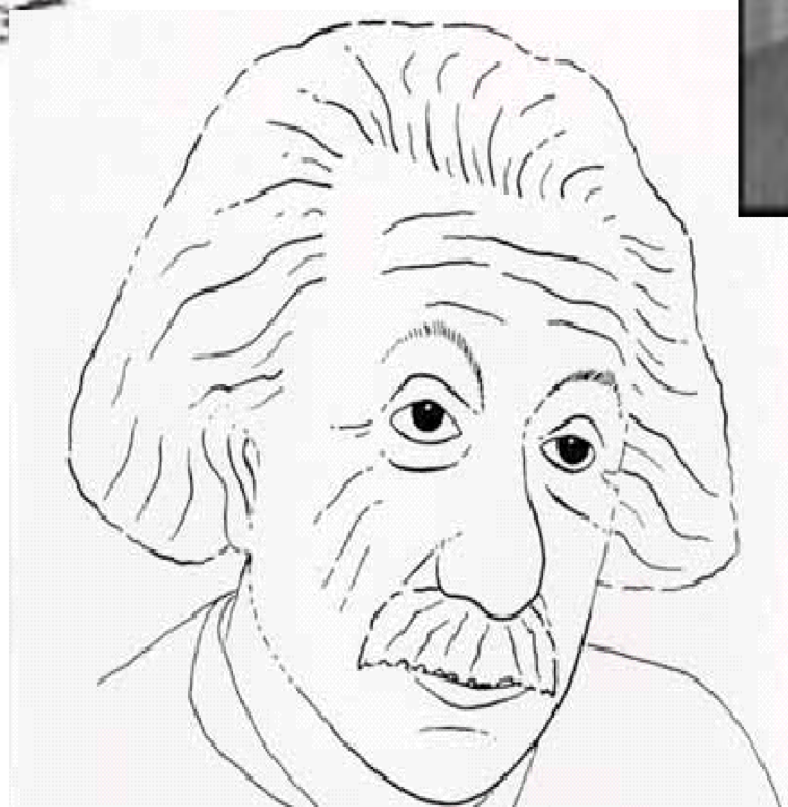
- Pervasiveness of video cameras
- At least half a million cameras in the city of London alone
- A single day a person could expect to be filmed 300 times while shopping in central London

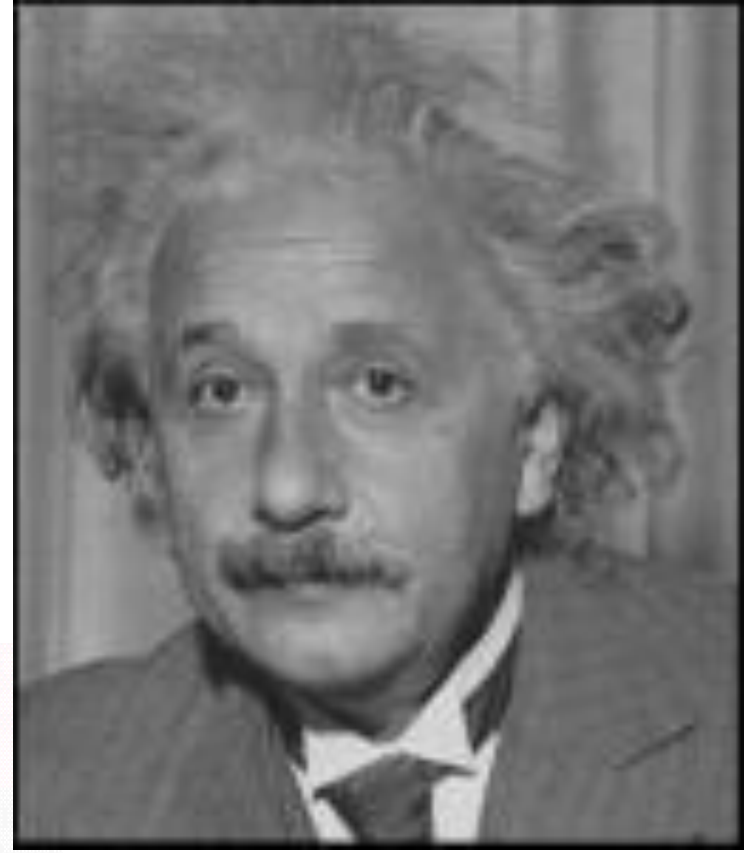
- « Everyone has the right to respect for his private life, his home and his correpsondence… »

  European Convention on Human Rights

Humans Vs. Machines

Humans recognize Cartoons and sketches





Machines (likely) fails

Humans Vs. Machines

Bill Clinton

Who is this guy?

Most computer algorithms are RST invariant

- Humans forget, machines do not!
- Humans need over 5 milliseconds to recognize a face
- Machines may recognize 5 faces in a millisecond

The role of beauty





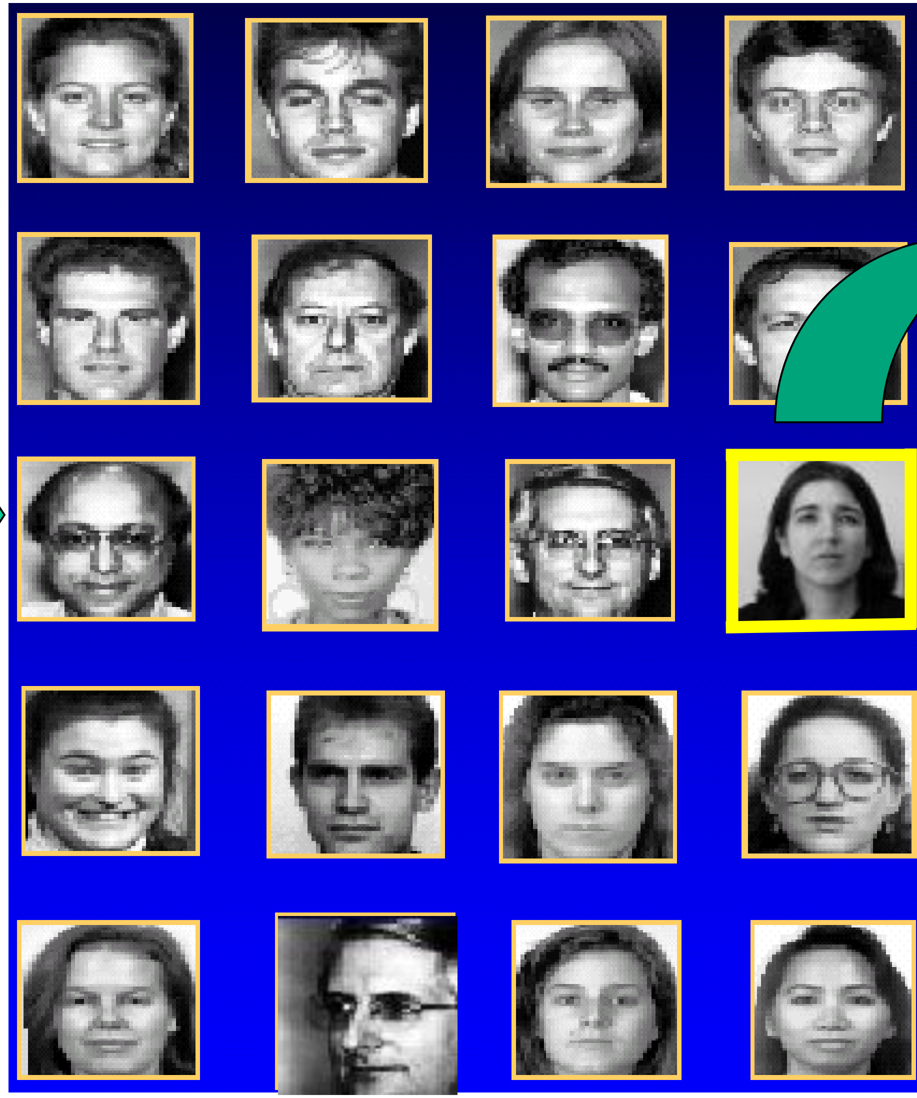
The “real” Miss Germany 2002 (Miss Berlin) has enough “features and patterns” to be recognized

The “virtual” Miss Germany, which was computed by blending together all contestants of the final round, is much more attractive

Source: http://www.beautycheck.de

The Challenge

Captured images of unknown person



Annotated Database

Miss Maria Smith

Queen Mary University of London
MMV Multimedia and Vision Research Group

The actual challenge….

Pictures taken form natural scenes

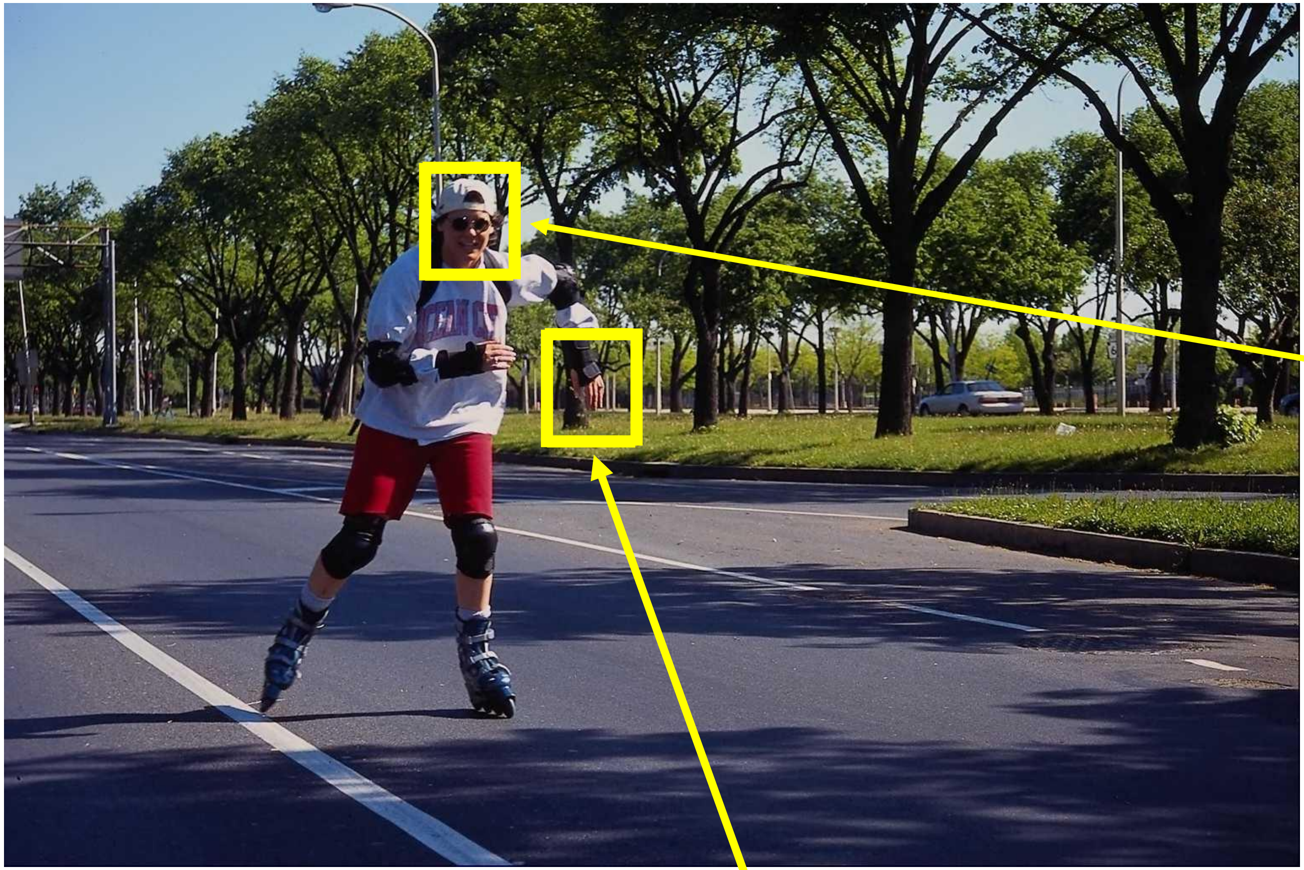
Detection

+

Recognition

Queen Mary University of London
MMV Multimedia and Vision Research Group

Detection



This is a face

This is NOT a face

Detection/Recognition (history)

- Kanade -First automated system- 1973
- Sirovich&Kirby –Principal component analysis- (1987)
- Turk&Pentland –Eigenfaces- (1991)
- Viola&Jones – AdaBoost and Haar Cascade (2004)
- Naruniec&Skarbek – Gabor Jets – (2007)
- Candès, Li and Ma – LowRank/Sparse Matrix separation (2013)
- Taigman, Yang, Ranzato –DeepFace – (2014)

Viola&Jones (very popular for detection)

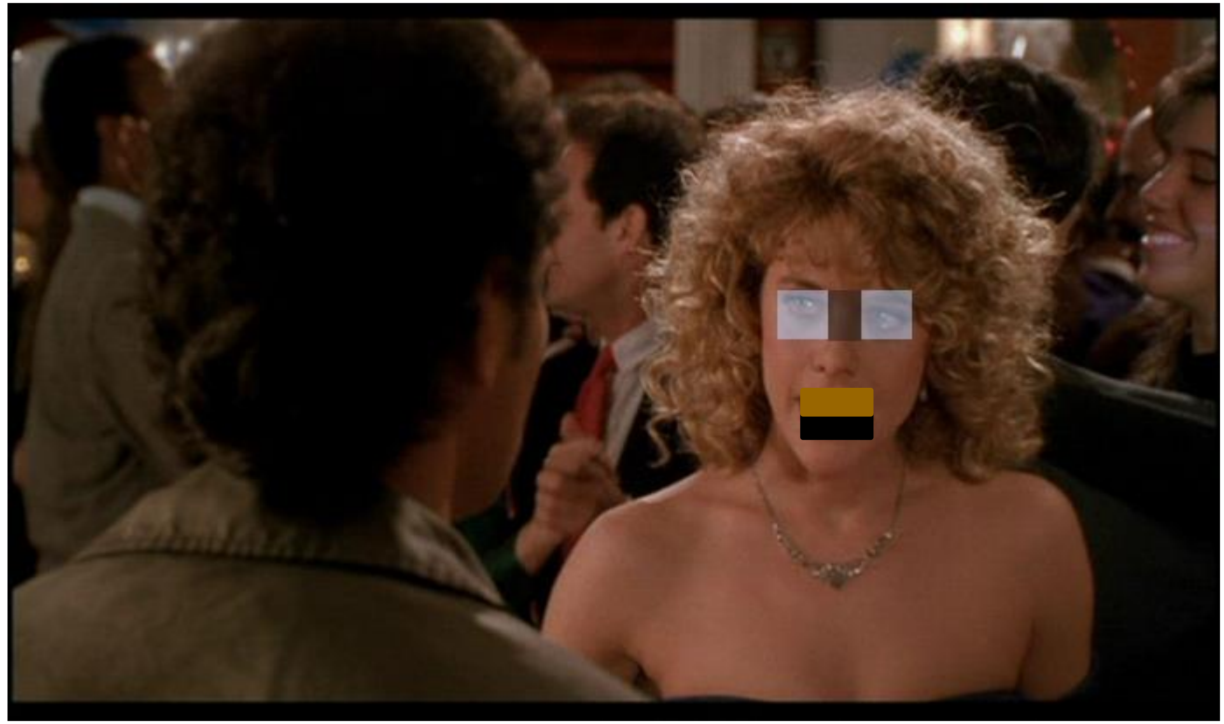
Idea:

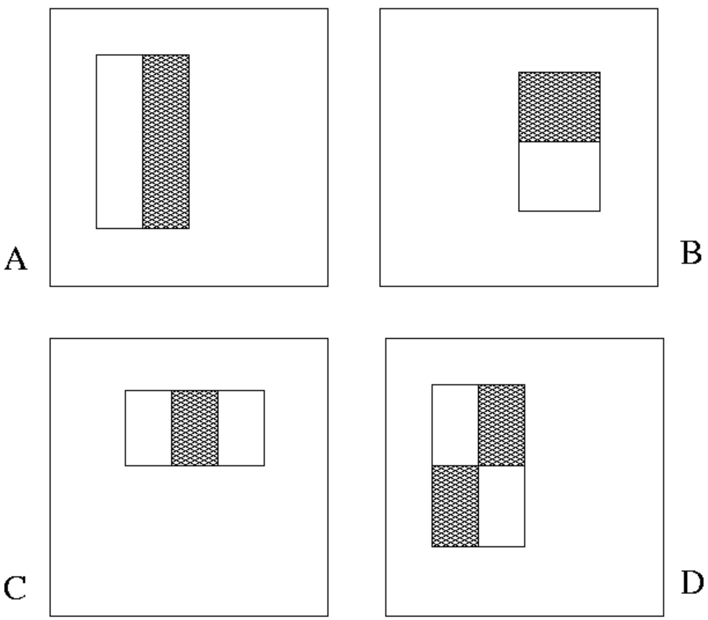
slide a window across image and evaluate a face model at every location

Viola&Jones

- Sliding window must evaluate tens of thousands of location/scale combinations
- Key ideas
  - *Evaluate image features using Haar-like filter responses*
  - *Integral images* for fast feature evaluation
  - *Boosting* for feature selection
  - *Attentional cascade* for fast rejection of non-face windows

Image Features





"Haar filter response"

Value = ∑ (pixels in white area)

– ∑ (pixels in black area)

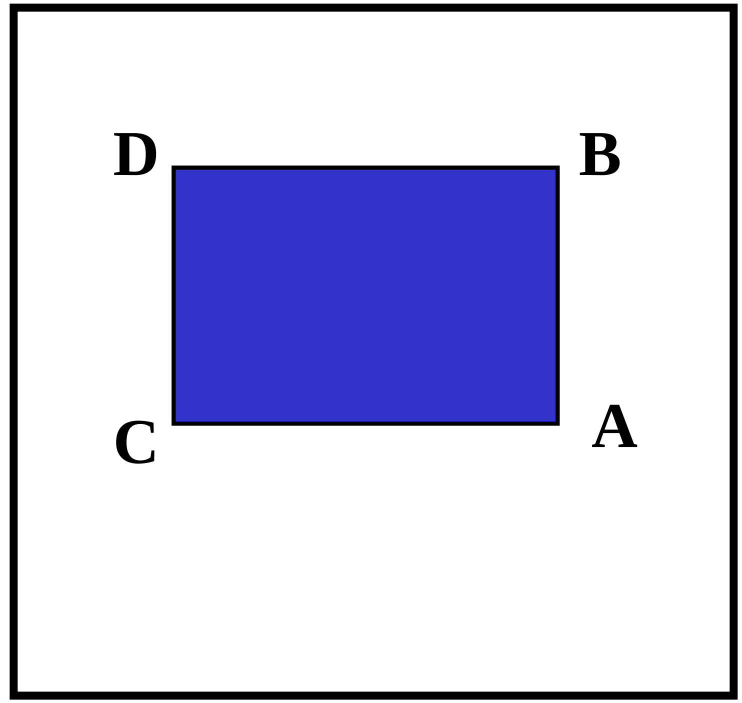
For a 24x24 detection region, the number of possible rectangle features is ~160,000! (computationally unfeasible)

Integral Images

- Let A,B,C,D be the values of the integral image at the corners of a rectangle
- Then the sum of original image values within the rectangle can be computed as:

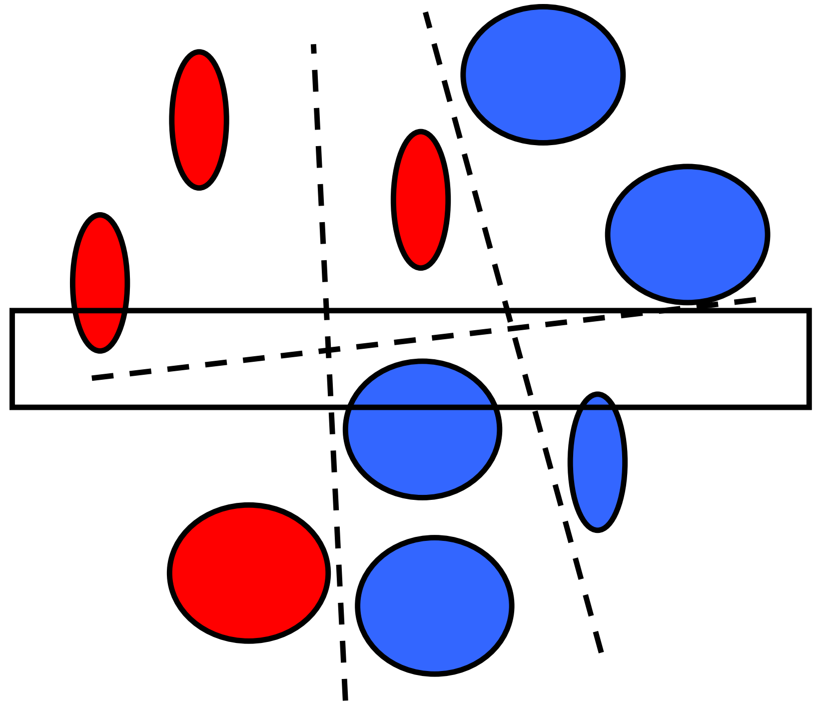
  Sum = A – B – C + D

- Only 3 additions are required for any size of rectangle!



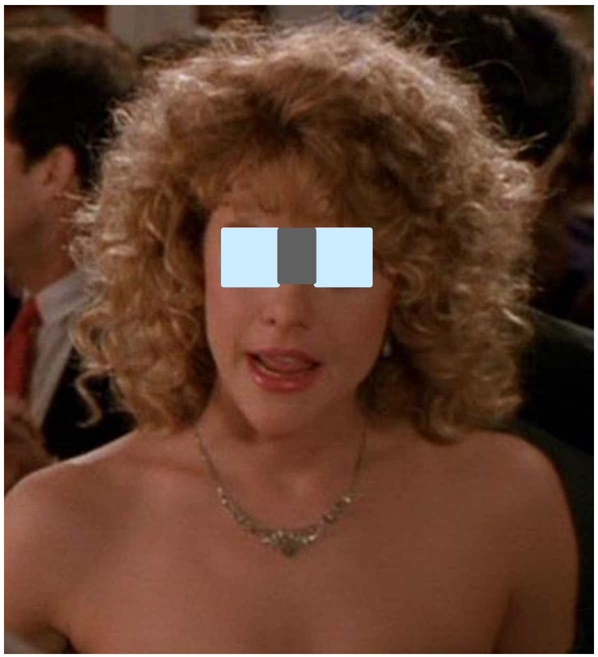
Boosting

Final classifier is a combination of weak classifiers



Boosting

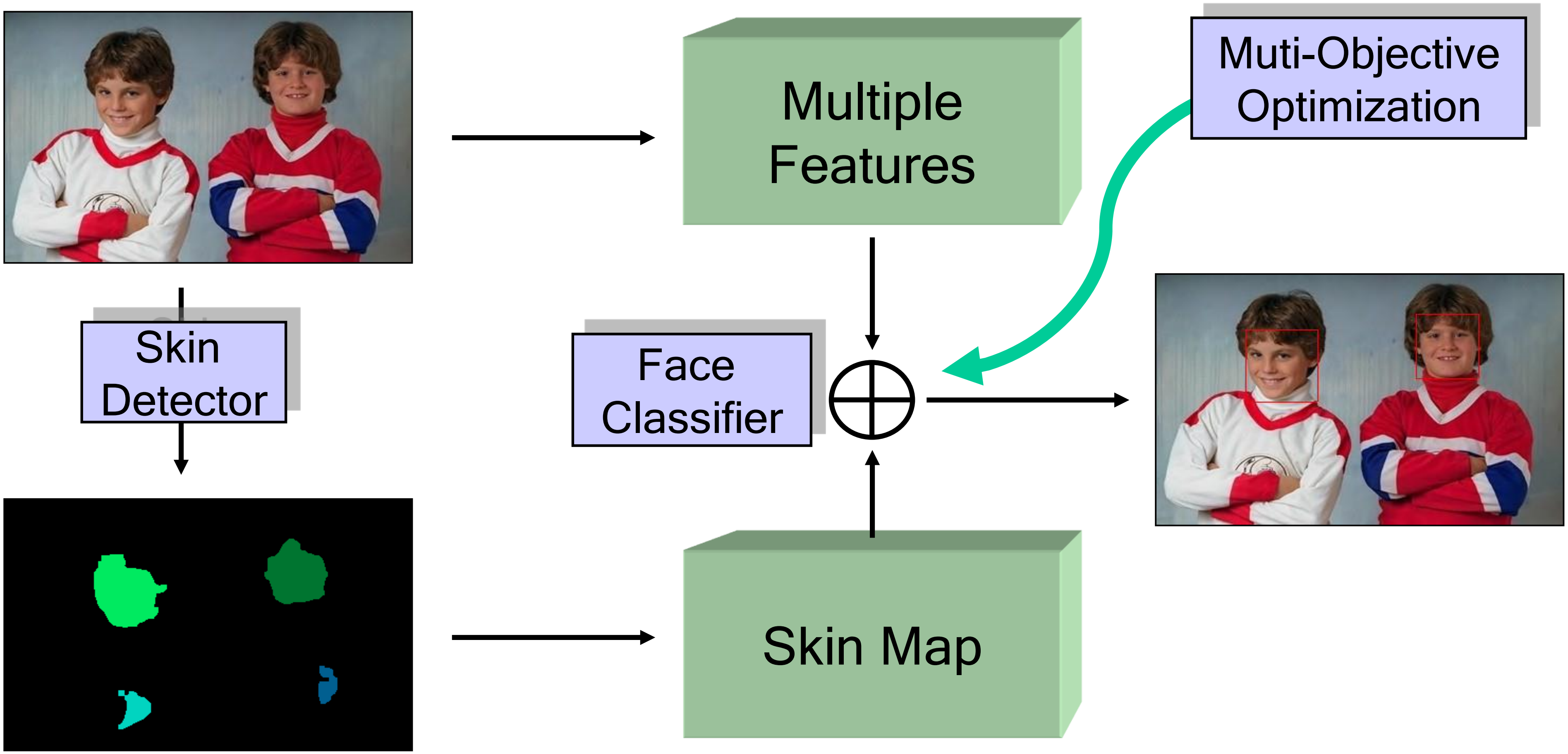
First two features selected by boosting:
This combination can yield 100% detection rate and 50% false positive rate


Other Approaches

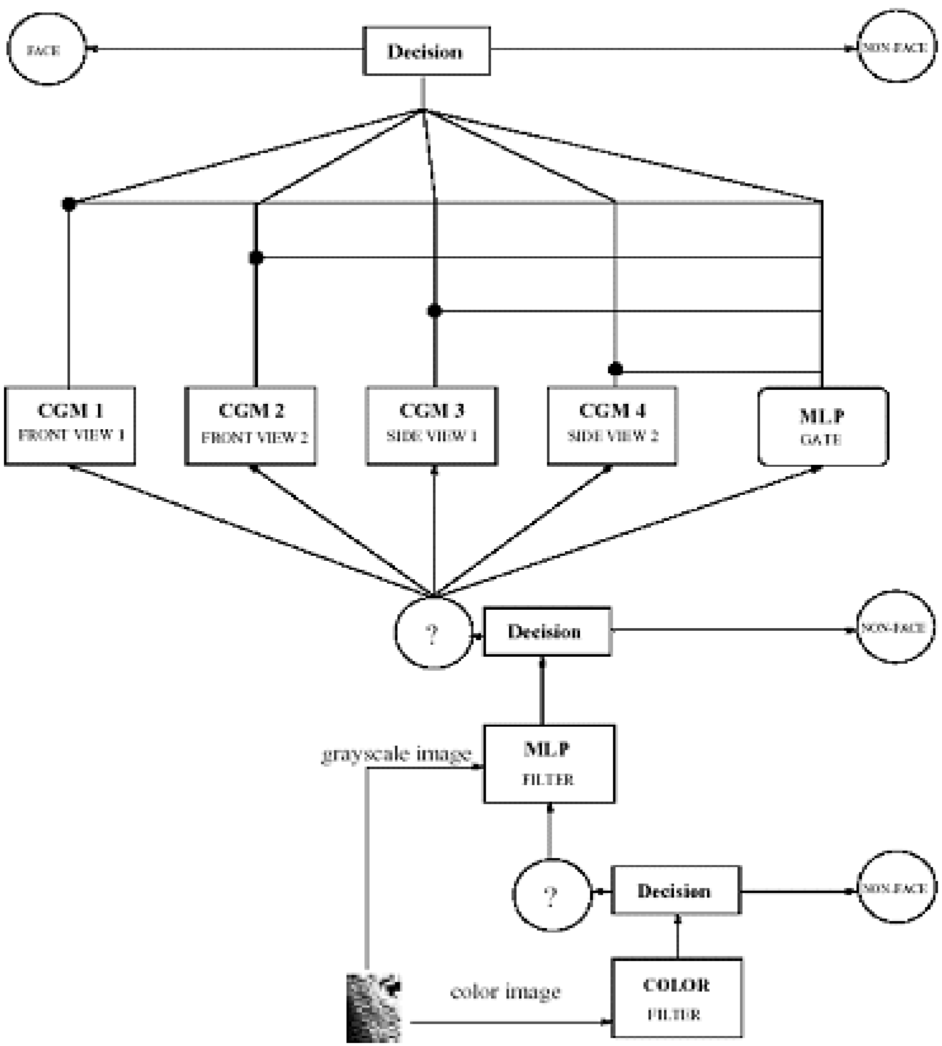
- Skin detection
  - Evaluation of different skin models in a suitable colour space and optimization
- Face classification
  - learning several descriptors suitable for human face: Edge distribution, Geometry, Texture distribution, Topology, etc.
  - Feature combination using MOO
  - face classification using support vector machines (SVM) (Djordjevic & Izquierdo 2009)

Overview QMUL Face detector

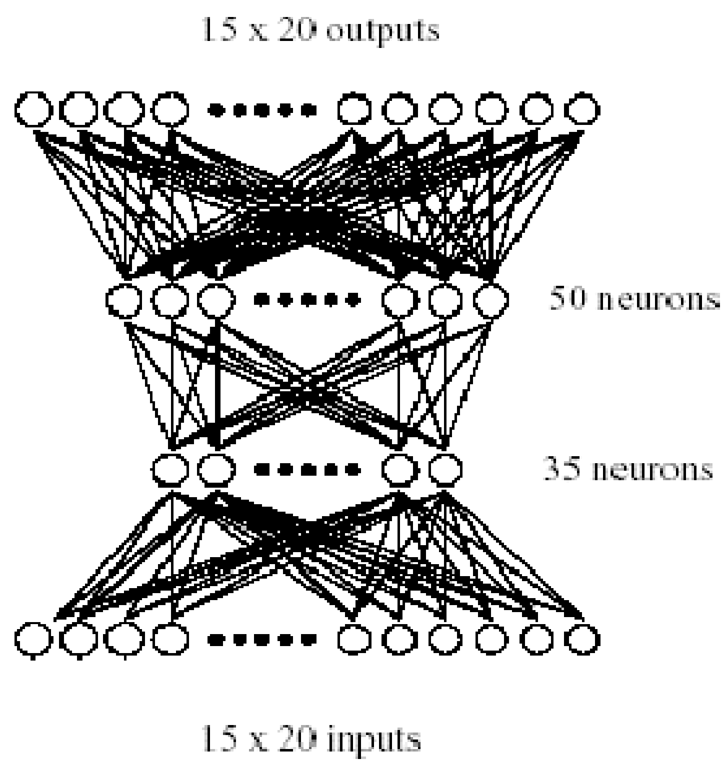


Queen Mary University of London
Multimedia and Vision Research Group

INRIA/FT Detector



Constrained Generative Model (CGM) Neural network based on the combination of several PCA-like classifiers



CGM Architecture

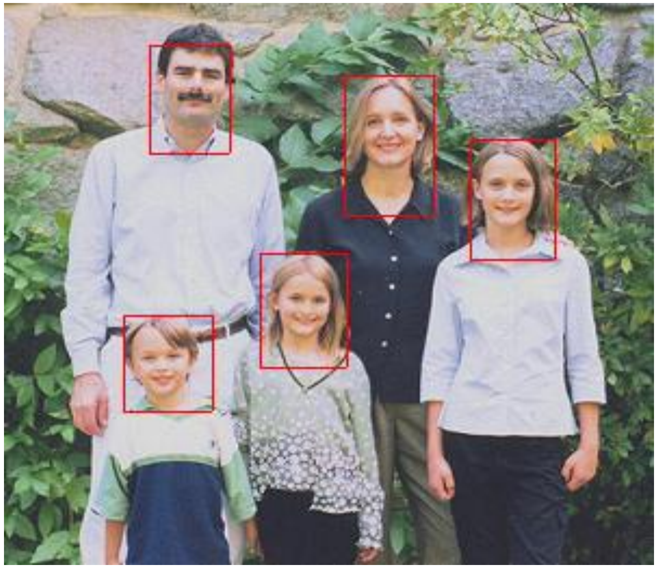
Database

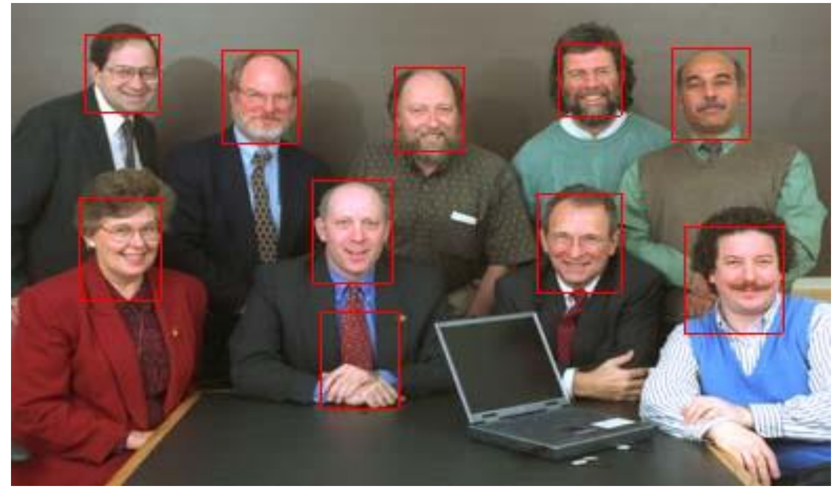
Setup of a test database with images coming from different sources

- CMU (130 images, 507 faces)
- Web (215 images, 499 faces)
- Cinema (162 images,276 faces)
- DiVAN (100 images, 104 faces)
- Yale (165 images, 165 faces)
- Stirling (207 images, 207 faces)
- Foreman (250 images, 250 faces)
- BioID (1521 images, 1525 faces)

A total of 3,533 test images with wide variability regarding face pose, occlusions, illumination conditions, facial expression, etc.

Face Detection (QMUL)





Queen Mary University of London MMV Multimedia and Vision Research Group

Face Detection (FT)

Queen Mary University of London
Multimedia and Vision Research Group

Results

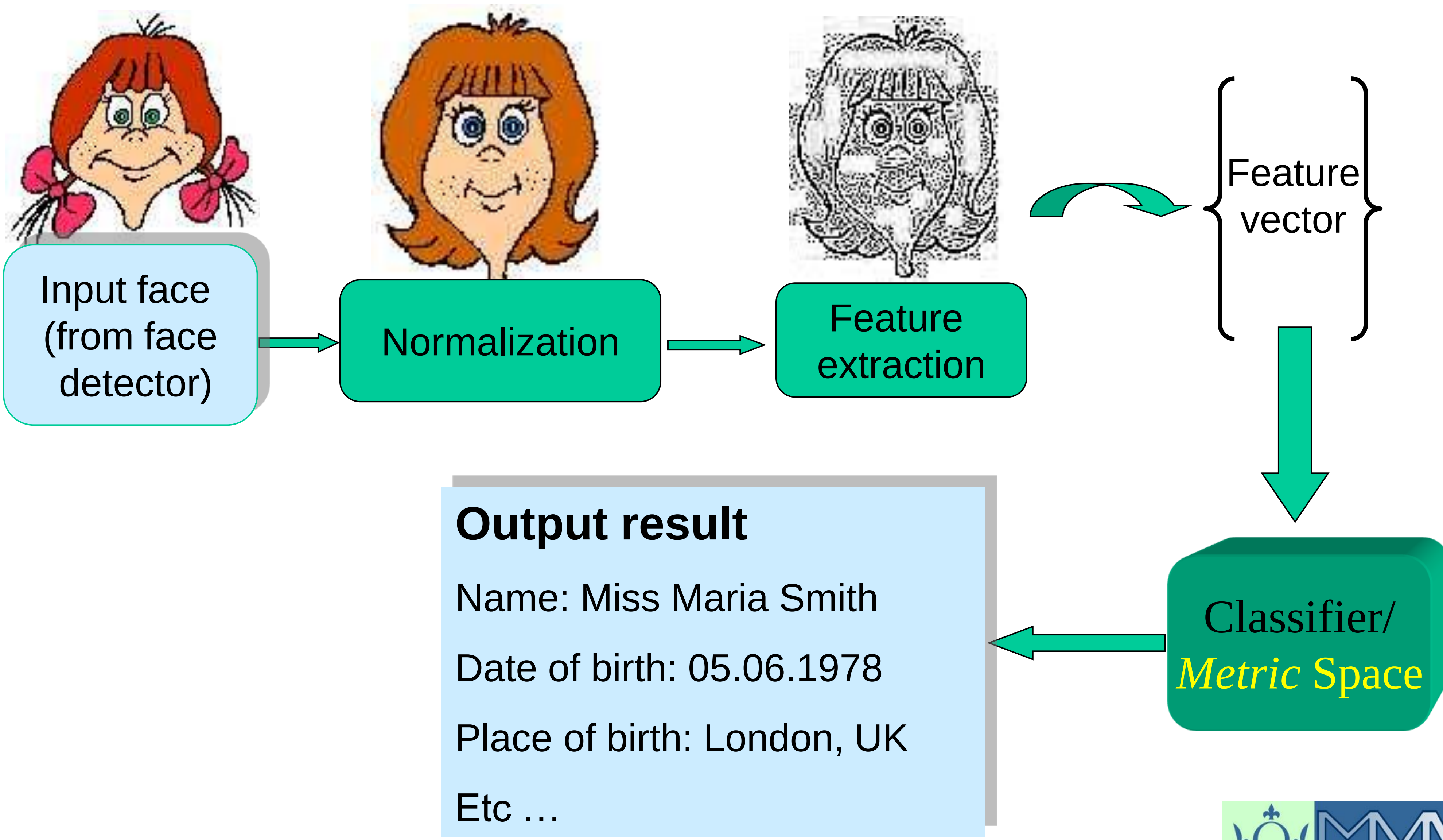
QMUL

- High accuracy

FT (CGN) is very fast

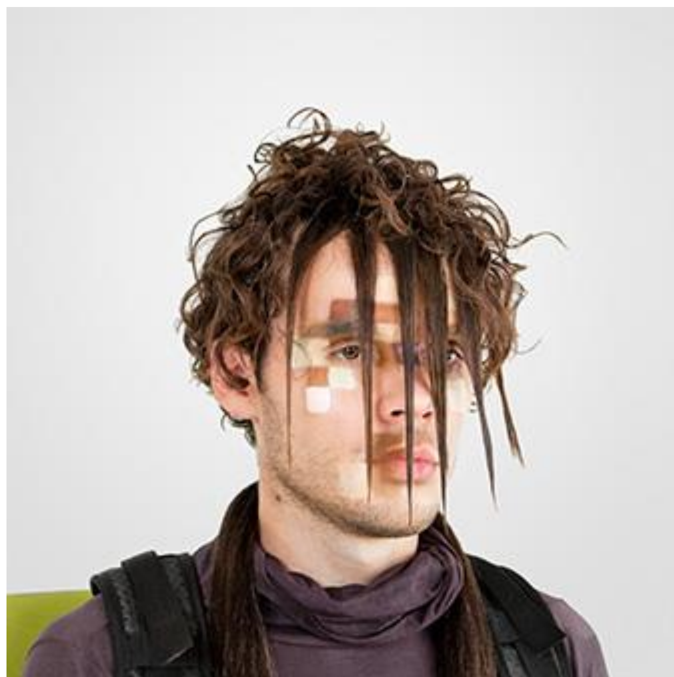
- Very efficient (330ms/image on a Pentium IV/3GHz)
- Low accuracy (sensitive to rotation)

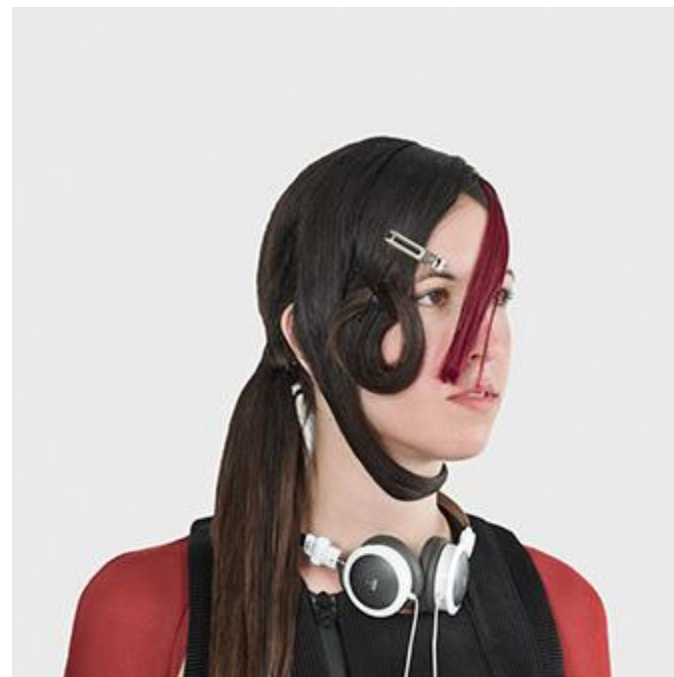
Recognition: Feature based matching



What is similarity?

CV Dazzle: Face & Anti-Face

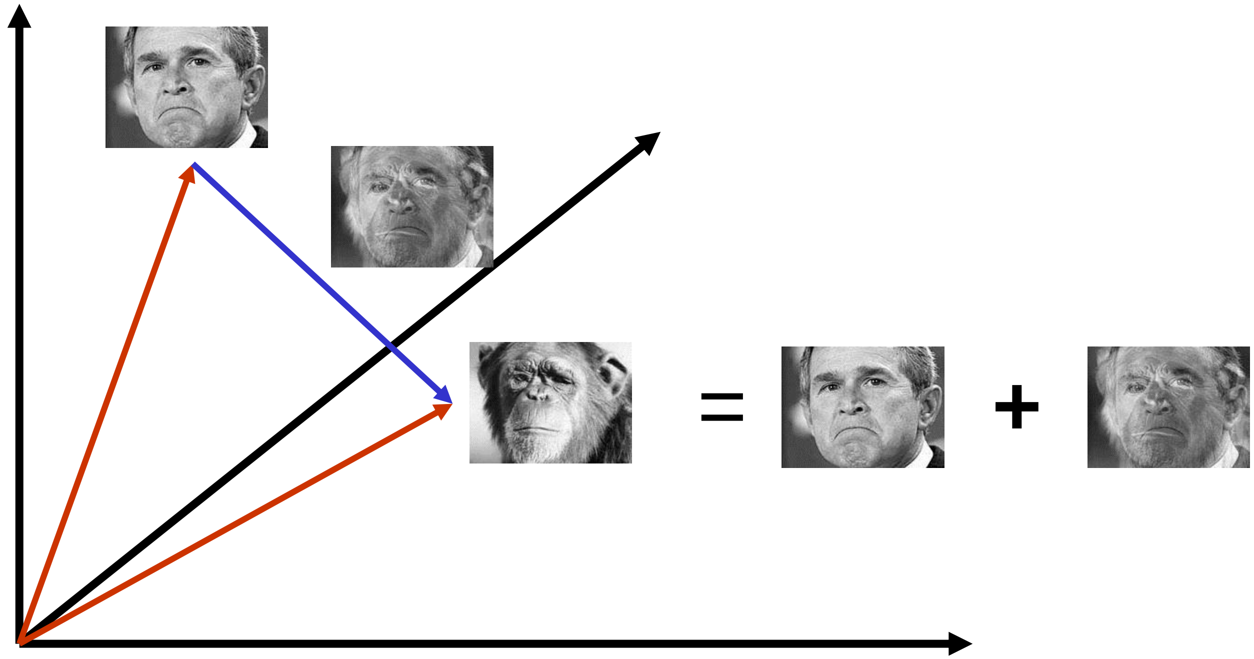




Recognition (selected approaches)

- Eigenfaces (classic approach)
- Using context in social networks
- Crowdsourcing and gamification
- “Indirect” identification
- Candès, Li and Ma – LowRank/Sparse Matrix separation (2013)

The Space of Faces



- An image is a point in a high dimensional space
  - An N x M image is a point in R^{NM}

Eigenfaces

Eigenfaces look somewhat like generic faces.

Projecting onto the Eigenfaces

- The eigenfaces $\mathbf{v_1}, \ldots, \mathbf{v_K}$ span the space of faces
  - A face is converted to eigenface coordinates by

$$\mathbf{x} \rightarrow (\underbrace{(\mathbf{x} - \overline{\mathbf{x}}) \cdot \mathbf{v}_1}_{a_1}, \underbrace{(\mathbf{x} - \overline{\mathbf{x}}) \cdot \mathbf{v}_2}_{a_2}, \ldots, \underbrace{(\mathbf{x} - \overline{\mathbf{x}}) \cdot \mathbf{v}_K}_{a_K})$$

$$\mathbf{x} \approx \overline{\mathbf{x}} + a_1\mathbf{v_1} + a_2\mathbf{v_2} + \ldots + a_K\mathbf{v_K}$$

$\mathbf{x}$

$a_1\mathbf{v}_1$ $a_2\mathbf{v}_2$ $a_3\mathbf{v}_3$ $a_4\mathbf{v}_4$ $a_5\mathbf{v}_5$ $a_6\mathbf{v}_6$ $a_7\mathbf{v}_7$ $a_8\mathbf{v}_8$

Recognition with Eigenfaces

Algorithm

1. Process the image database (set of images with labels)
   - Run PCA—compute eigenfaces
   - Calculate the K coefficients for each image
2. Given a new image (to be recognized) **x**, calculate K coefficients

$$\mathbf{x} \rightarrow (a_1, a_2, \ldots, a_K)$$

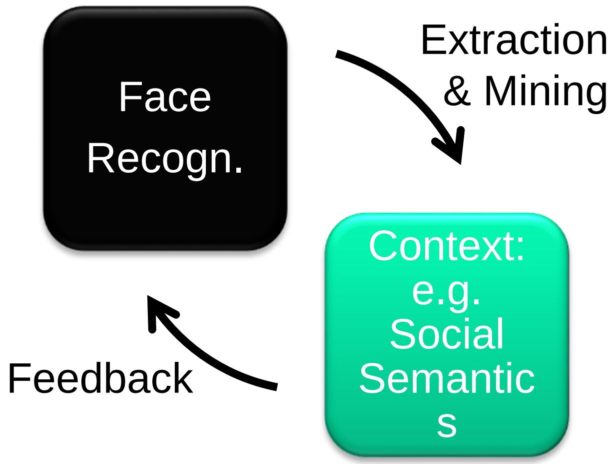
3. Detect the closest labeled face in the database

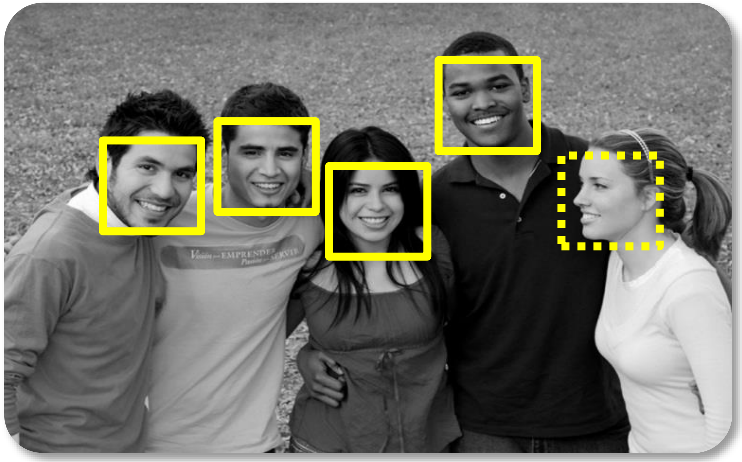
Joint People Recognition across Photo Collections using BN

Markus Brenner, Ebroul Izquierdo

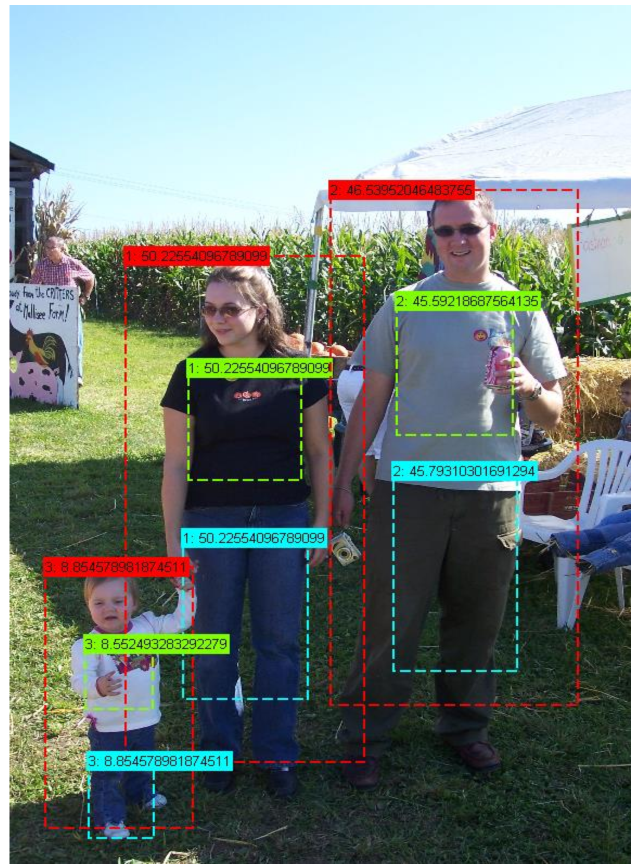
Queen Mary University of London
MMV Multimedia and Vision Research Group

Improve Face Recognition through Context





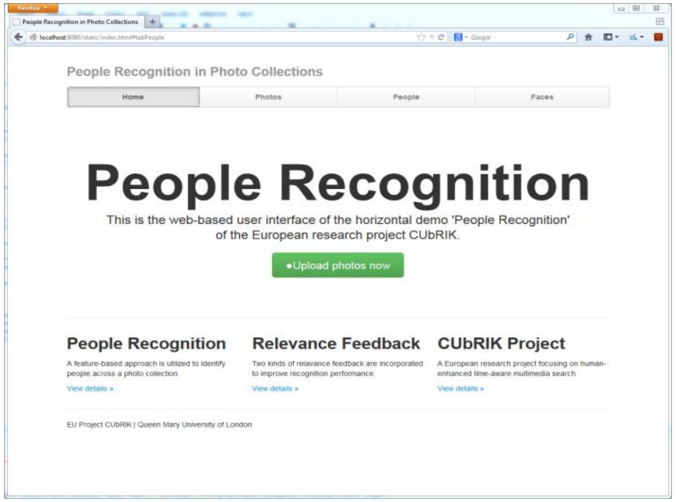
Detecting Body, Extracting Clothing



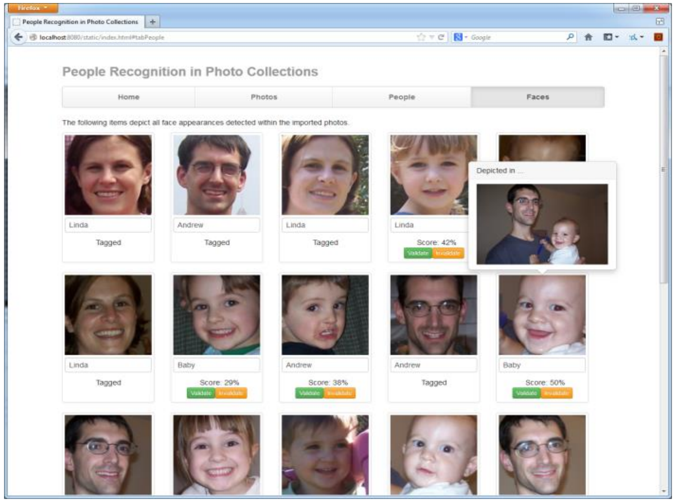
Queen Mary University of London

Multimedia and Vision Research Group

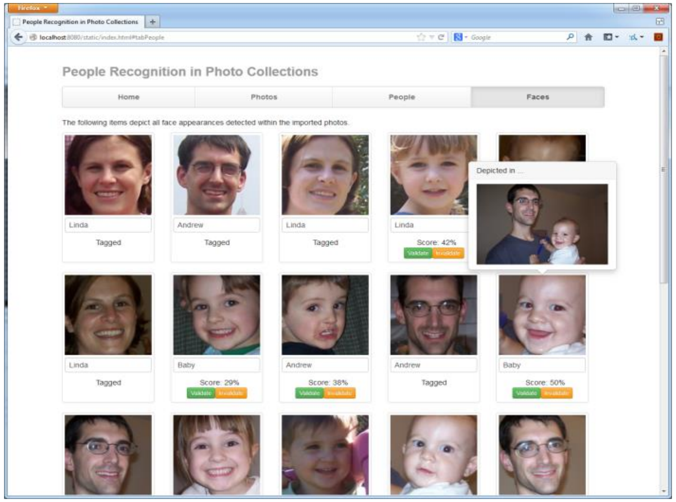
Adding Crowdsourcing



Dynamic (AJAX)



Gamification



RESTful API

A useful and “more practical” approach

In a security scenario

- Identify faces “indirectly” by identifying related DROPs (Distinctive Regions or Patters)
- Reduces to a “query by example search”
- DROP tracing and tracking should be robust in
  - low frame rate video sequences
  - the presence of camera-shake
  - blurred images
  - illumination changes caused by flash-lighting such as emergency vehicle lighting, fire and flash photography

Example - a guy with an orange zip

Queen Mary University of London
Multimedia and Vision Research Group


CH2 720x288
L-PLAY 04


13/12/06 13:43:53 08

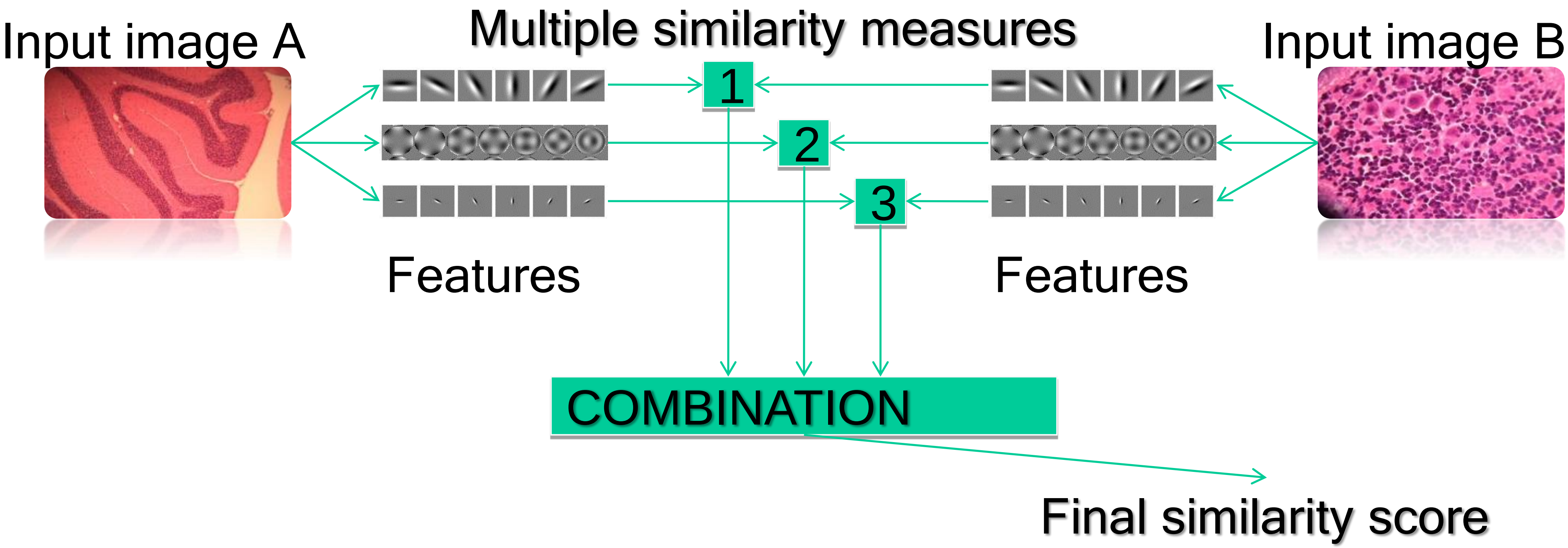
Example - a guy with an orange zip

Face recognition by DROP tracing

Good example for Identification & Prosecution.

Robust DROP tracing

Improve discrimination power by combining multiple low-level features



Q. Zhang and E. Izquierdo, "Adaptive Salient Image Retrieval in Multi-Feature Space," Signal Processing: Image Communication, 2007.

Problems in feature combination

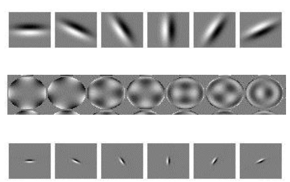
Use of simple vector concatenations

Problems:

- Most feature spaces are non-linear
- Different feature spaces => different distance metrics
- Feature concatenation => distance estimated by uniform metrics

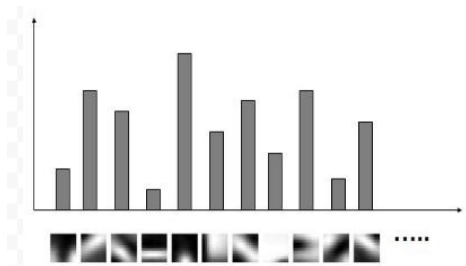
3 feature vectors

- Gabor textures
- Tamura textures
- Zernike moments

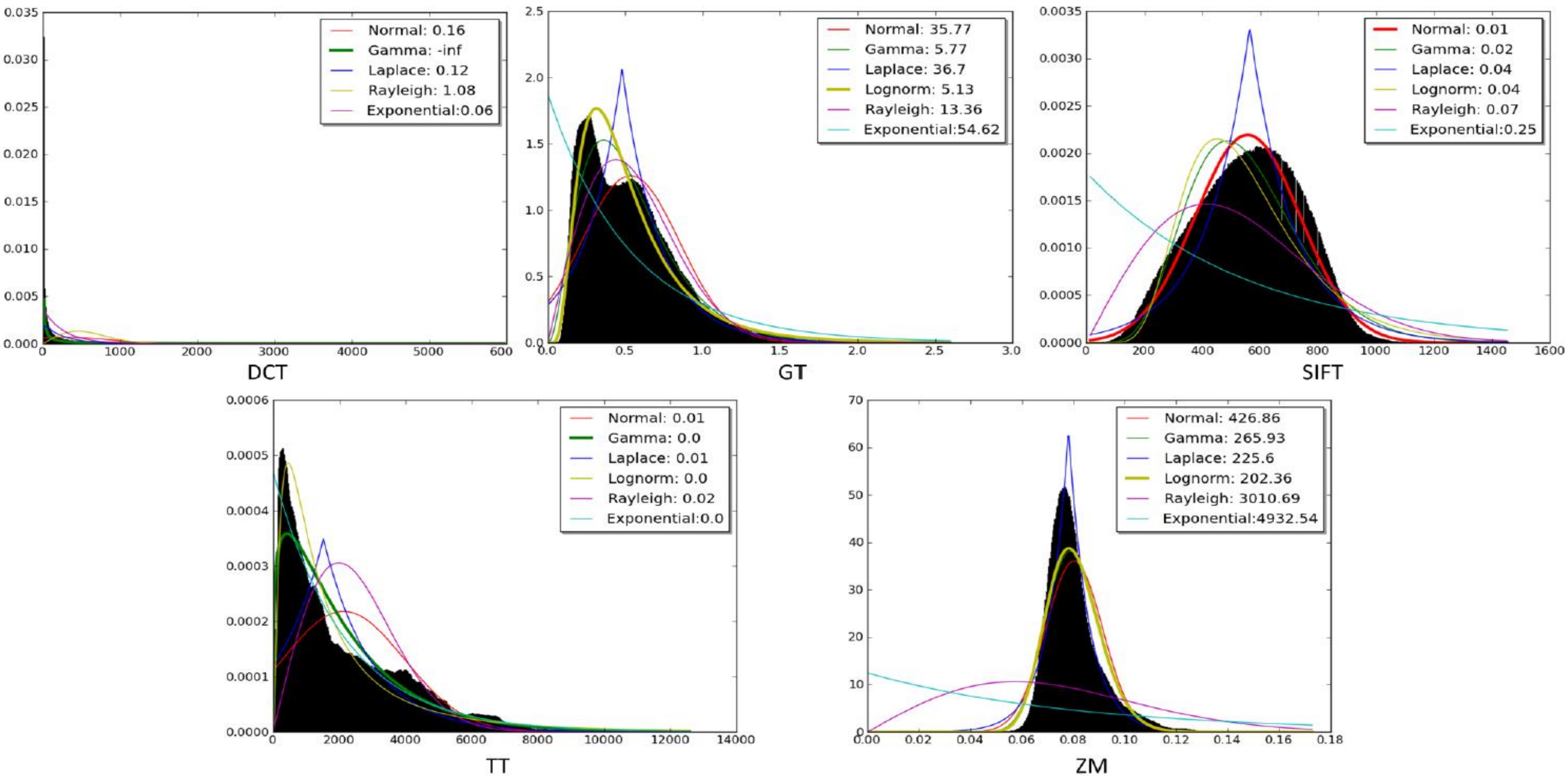


2 feature histograms

- SIFT-based Bag-of-features
- DCT-based Bag-of-features



Feature Normalization



Each feature follows its own distance distribution

Feature Combination Metric

Feature combination metric:

$$D(I_1, I_2) = \sum_{j \in \mathbf{F}} \omega_j d(v_{1,j}, v_{2,j})$$

Weighting factors:

$$\{\omega_j, j \in \mathbf{F}\}$$

How to find optimized weighting factors?

Multi-objective optimisation

Multiple objective functions

Conflicting interest

General optimum

$$D_1 = \sum_{j=1}^{q} \omega_j d_{1,j}$$

$$D_2 = \sum_{j=1}^{q} \omega_j d_{2,j}$$

$$D_3 = \sum_{j=1}^{q} \omega_j d_{3,j}$$

$$\cdots$$

Objective functions

Multiple objective optimisation

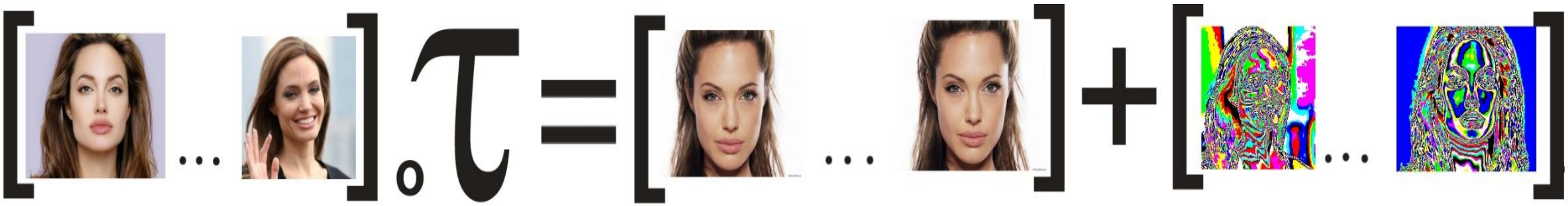
Training set

Q. Zhang and E. Izquierdo, “Combining low-level features for semantic inference in image retrieval,” EURASIP Journal on Advances in Signal Processing, 2009

Face alignment as critical component for recognition

Low-rank and sparse matrix decomposition

Sound model with excellent results



Corrupted/ Misaligned Faces

Aligned Low-Rank Images

Sparse Component

Low-rank and sparse matrix decomposition

- Consider different faces as perturbations of a normalized/aligned face
- Aligned: A low rank approximation of the matrix X
- Perturbation: A sparse matrix
- Mathematical formulation: X = L + S

  where L and S solve the minimization problem:

$$\min_{L,S} \left\{ rank(L) + \lambda \|S\|_0 \right\} \quad st \quad L + S = X$$

- Main limitations: Nonconvexity of the discrete-valued functions matrix rank and l_0-norm,

A novel approach: RPCA **(Candes et al., 2009)**

- This separation problem seems impossible to solve
- Very surprisingly, not only can this problem be solved, but it can be solved by convex optimization using PCP to solve

$$\begin{array}{ll} \text{minimize} & \|L\|_* + \lambda\|S\|_1 \\ \text{subject to} & L + S = M \end{array}$$

- It works even if the rank of L grows almost linearly in the dimension of the matrix and the errors in S are up to a constant fraction of all entries.

Theorem 1.1 *Suppose L_0 is $n \times n$, obeys (1.2)–(1.3), and that the support set of S_0 is uniformly distributed among all sets of cardinality m. Then there is a numerical constant c such that with probability at least $1 - cn^{-10}$ (over the choice of support of S_0), Principal Component Pursuit (1.1) with $\lambda = 1/\sqrt{n}$ is* exact, *i.e. $\hat{L} = L_0$ and $\hat{S} = S_0$, provided that*

$$\operatorname{rank}(L_0) \le \rho_r n\, \mu^{-1}(\log n)^{-2} \quad \text{and} \quad m \le \rho_s n^2. \tag{1.4}$$

Above, ρ_r and ρ_s are positive numerical constants. In the general rectangular case where L_0 is $n_1 \times n_2$, PCP with $\lambda = 1/\sqrt{n_{(1)}}$ succeeds with probability at least $1 - cn_{(1)}^{-10}$, provided that $\operatorname{rank}(L_0) \le \rho_r n_{(2)}\, \mu^{-1}(\log n_{(1)})^{-2}$ *and $m \le \rho_s n_1 n_2$.*

τ-Decomposition for face alignment

Guerra, Izquierdo and Erfanian (2014)

Let $F_1, F_2, ... F_n$ be different “distorted photos” from the same person, then there exist transformations such that $\tau_1, \tau_2, ... \tau_n$

$$F_1 o \tau_1, F_2 o \tau_2, ... F_n o \tau_n$$

represent the “base face”

How are the transformation parameters calculated?

The parameters $\tau_1, \tau_2, ... \tau_n$ should guarantee a Low Rank +Sparse decomposition of the matrix X

τ-Decomposition for face alignment

Guerra, Izquierdo and Erfanian (2014)

- We consider affine transformations, i.e., τ_i is a 6-vector, *i=1,…,n*
- The optimization problem includes τ_i as variable.
- Using a Lagrange formulation, we solve:

$$\min_{L,S,\tau} \left\{ \| X o \tau - L - S \|_F^2 + \lambda \|S\|_1 \right\}$$

- by an alternating iterative strategy:

(A) $$\tau^t = \arg\min_{\tau} \| X o \tau - L^{t-1} - S^{t-1} \|_F^2$$

(B) $$L^t = \arg \min_{rank(L) \leq r} \| X o \tau^t - L - S^{t-1} \|_F^2$$

(C) $$S^t = \arg\min_{S} \| X o \tau^t - L^t - S \|_F^2 + \lambda \|S\|_1$$

Queen Mary University of London

MMV Multimedia and Vision Research Group

τ-Decomposition for face alignment

Guerra, Izquierdo and Erfanian (2014)

Theorem A:

The minimization problem (A) is equivalent to a weighted least squares problem and thus it has a closed-form solution

Theorem B:

The minimization problem (B) has a the following closed-form solution

$$L^t = \sum_{i=1}^{r} \sigma_i U_i V_i^T \ , svd(X \circ \tau^t - S^{t-1}) = U \Sigma V^T$$

Note: Many SVD required!!

τ-Decomposition for face alignment

Guerra, Izquierdo and Erfanian (2014)

Solving (C):

- The matrix S^t is updated using the parameter λ as a threshold in

$$X \circ \sigma^t - L^t$$

- i.e, the elements of $X \circ \sigma^t - L^t$ less or equal than λ are considered to be zero

τ-Decomposition for face alignment

Guerra, Izquierdo and Erfanian (2014)

Convergence Theorem (*):

For a given λ, the sequence of values of the objective function

$$\left\| X \circ \tau^t - L^t - S^t \right\|_F^2 + \lambda \left\| S^t \right\|_1 , \; t = 1,2,\ldots$$

produced by the iterative process (A, B, C) is monotonically decreasing.

A "SVD-free" result

Izquierdo, Erfanian and Guerra (2014)

Let the rank of a low rank matrix L be 1 up to a given transform

Theorem B1: The solution of (b) is given by

$$l_i = \frac{1}{n}\sum_{j=1}^{n} E_{ij} \ , i = 1,...m$$

where

$$E = X o \tau^t - S^{t-1}$$

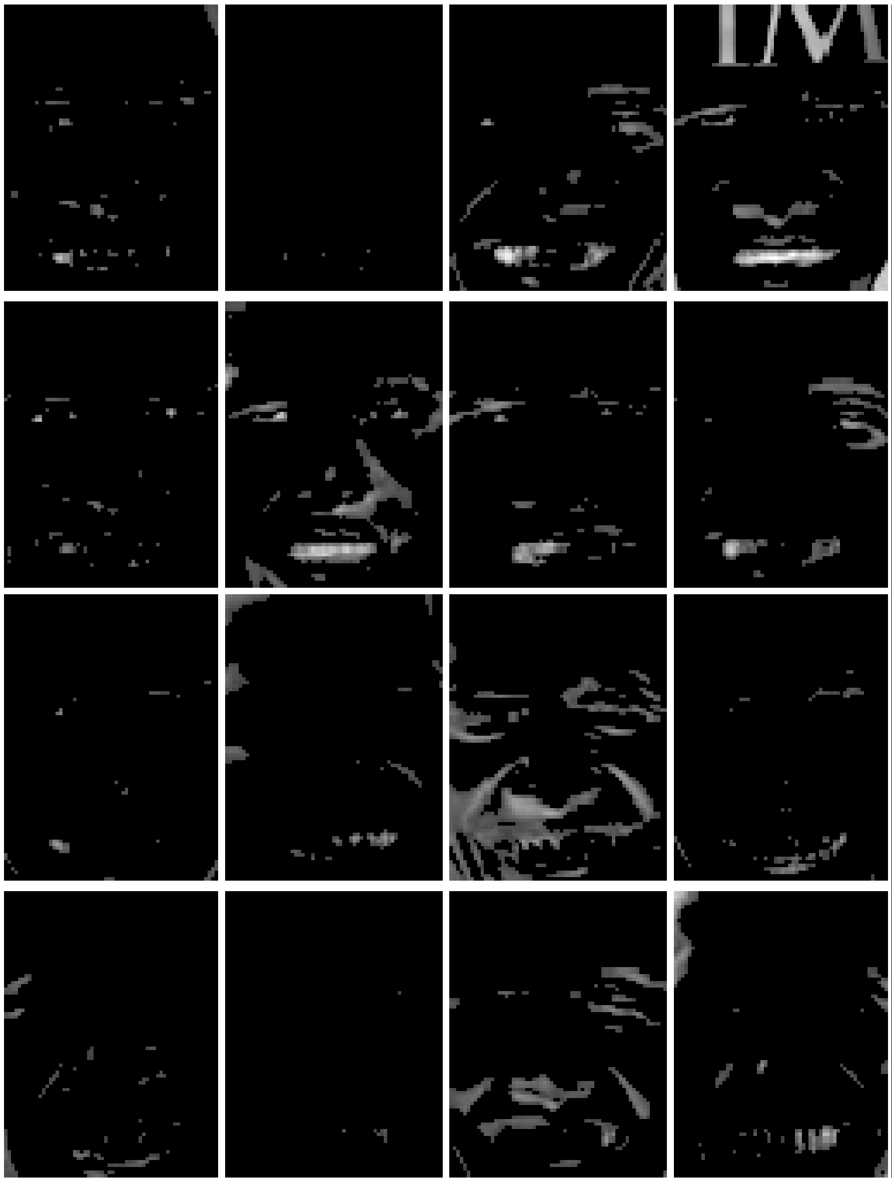

HOUSE

τ-Decomposition for face alignment

Guerra, Izquierdo and Erfanian (2014)

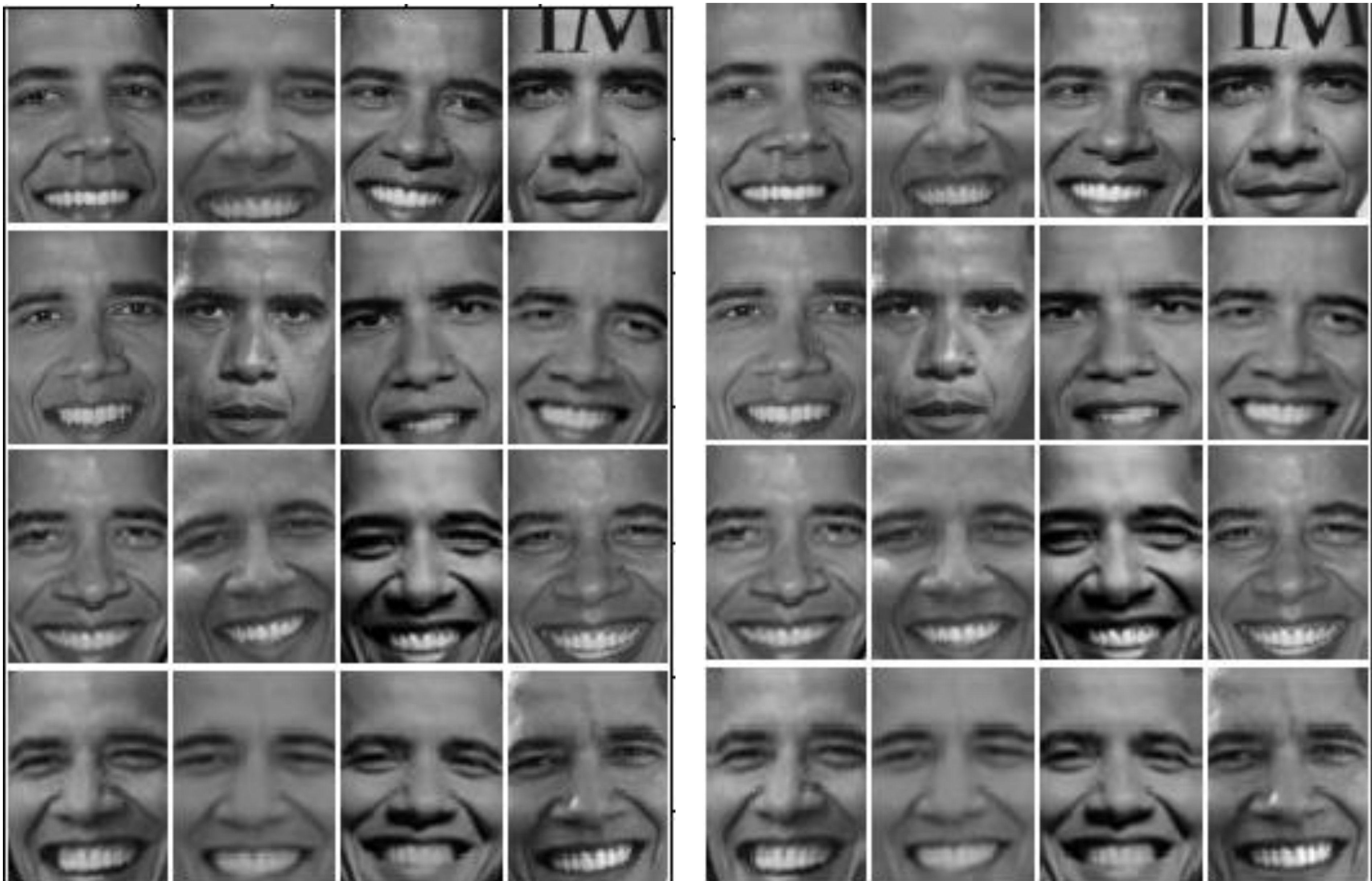
τ-Decomposition for face alignment

Guerra, Izquierdo and Erfanian (2014)



τ-Decomposition for face alignment

Guerra, Izquierdo and Erfanian (2014)



Where are we now?

Medium complexity

Good solutions available

Where are we now?

High complexity: Crowds, many objects –Substantial R&D needed-

Where are we now?

CCTV content

–Substantial R&D needed-

Looking out of the box

Perhaps learning from super-recognizers

Queen Mary University of London
Multimedia and Vision Research Group

Can you recognize this person?

Can a machine do it?

Queen Mary University of London

Multimedia and Vision Research Group

Suspect identified by "super recogniser" PC Gary Collins. Convicted of - Violent Disorder, Robbery, Burglary, Arson in London 2011 riots. Sentence - 6 years in prison

Queen Mary University of London

MMV Multimedia and Vision Research Group


ITV 1
26/09/2013

Acknowledgments and thanks to the VideoSense/Lasie project and other partners, including:

- Scotland Yard (The MET London)
- The University of Greenwich
- The EU funded Lasie Project

LASIE
LARGE SCALE INFORMATION
EXPLOITATION OF FORENSIC DATA

M. Neville

J. Davis

V. Guerra

D. Djorjevic

Q. Zhang

Queen Mary University of London

MMV Multimedia and Vision Research Group

Where are we now?

Thanks

Queen Mary University of London

MMV Multimedia and Vision Research Group