Face Detection and Recognition
-From forensics and biometrics to social semantics-

Ebroul Izquierdo
Multimedia and Vision Group
Queen Mary, University of London

VIDESENSE 2014 Winter School on Multimodal and Privacy-Respecting Video Analytics
Pont Saint Martin, Italy, November 26th 2014
Road Map

• Motivation
• Face detection and classifiers
• Face recognition
• Eigenfaces
• Social aspects (crowdsourcing)
• The human factor
• Latest from linear algebra
• Concluding remarks
Motivation and context

• Serious societal problems derived from heightened vandalism
• The fatal consequences of terrorist activities
• The fact that advanced computer vision and multimedia signal processing can provide key tools to solve or at least lighten such societal problems
• Automated recognition of people and their most fundamental characteristics or traits plays a crucial role.
• Face recognition is one of the main biometric tools in automated authentication.
Motivation and context

• Face recognition is one of the main biometric used in security applications

Furthermore

• Key in visual information retrieval
• It embraces many tasks: from simple classification or clustering to conceptual recognition

Let’s looks at the features of face recognition!
The good

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

• Biometrics
• Authentication
  – No more password and PIN
  – Cannot be stolen
  – Cannot leave it at home

“Sorry about the odor. I have all my passwords tattooed between my toes.”
The bad

• Very difficult task
• Important commercial applications requires high accuracy
• Derived technology can be intrusive (specially in surveillance applications)
The ugly
It has severe privacy and ethical implications

The pervasive use of video cameras is part of daily life in most cities of the world. As an example, the London underground network installed its first cameras, using black-and-white film, back in 1961 at the Holborn station in central London. By the 1990s, every single station had a surveillance system. They are now all linked to three control rooms on London's outskirts. There are at least half a million cameras in the city of London and in a single day a person could expect to be filmed 300 times.
The better

Humans Vs. Machines

• Humans recognize Cartoons
• Machines (likely) fails
The better Humans Vs. Machines

- Humans recognise sketches
- Machine (likely) fails
The better

Humans Vs. Machines

Who is this guy?  Bill Clinton

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 beauty

Some of the most beautiful faces (from Germany)

Miss Baden-Württemberg, Bavaria, North-Rhine/Westphalia, South Germany, Thuringia
The beauty

The most beautiful face
It is virtual (computer generated)
... and can not be “recognized”
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: Beauty check, http://www.beautycheck.de
The Challenge

Captured images of unknown person

Annotated Database

Miss Maria Smith
The actual challenge....

Pictures taken from natural scenes are very complex.
Consequently…

Two main tasks need to be tackled
• Detection (where is the face)
• Matching (actual recognition)
Road Map

• Motivation

Face detection and classifiers
• Face recognition
• Eigenfaces
• Social aspects (crowdsourcing)
• The human factor
• Latest from linear algebra
• Concluding remarks
Detection

This is a face

This is NOT a face

This is a face
Accurate face detection (QMUL face detector)

- Skin detection (Liu & Izquierdo 2005)
  - Evaluation of different skin models in a suitable colour space and optimization
- Face classification (Liu & Izquierdo 2008)
  - Learning several descriptors suitable for human face: Edge distribution, Geometry, Texture distribution, Topology, etc
  - Face classification using support vector machine (SVM) (Djordjevic & Izquierdo 2008)
Accurate face detection (QMUL face detector)

- Face classification (Zhang & Izquierdo 2008/2009)
  - learning descriptors
  - face classification by (SVM)
  - Multi-objective optimization of descriptor combination in high dimensional metric spaces

It achieves higher detection accuracy and fast response
Overview QMUL Face detector

Skin Detector

Edge Map

Face Classifier

Skin Map
Road Map

- Motivation
- Face extraction
- SVM as face classifiers
  - Multi-objective optimization of feature spaces
  - Face recognition
  - Eigenfaces
- Concluding remarks
SVM as Face classifier

- Binary classification is viewed as the task of separating two classes in feature space:

$$w^T x + b > 0$$

$$w^T x + b < 0$$

$$f(x) = \text{sign}(w^T x + b)$$

Linear Separators

• Which of the linear separators is optimal?
Classification Margin

- Distance from example $x_i$ to the separator is

$$r = \frac{\mathbf{w}^T \mathbf{x}_i + b}{\|\mathbf{w}\|}$$

- Examples closest to the hyperplane are support vectors.

- Margin $\rho$ of the separator is the distance between support vectors.
Maximum Margin Classification

• Maximizing the margin is good according to intuition and PAC (probably approximated correct learning) theory.
• Implies that only support vectors matter; other training examples are ignorable.
Non-linear SVMs: Feature spaces

- The original feature space can always be mapped to some higher-dimensional feature space where the training set is separable:

\[ \Phi: \mathbf{x} \rightarrow \varphi(\mathbf{x}) \]
The “Kernel Trick”

• The linear classifier relies on inner product between vectors $K(x_i, x_j) = x_i^T x_j$

• If every datapoint is mapped into high-dimensional space via a transformation $\Phi: x \rightarrow \varphi(x)$, the inner product becomes:

$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$$

• A kernel function is a function that is equivalent to an inner product in some feature space.
Examples of Kernel Functions

- Linear: $K(x_i, x_j) = x_i^T x_j$
  - Mapping $\Phi$: $x \rightarrow \varphi(x)$, where $\varphi(x)$ is the identity

- Polynomial of degree $p$: $K(x_i, x_j) = (1 + x_i^T x_j)^p$
  - Mapping $\Phi$: $x \rightarrow \varphi(x)$, where $\varphi(x)$ has dimensions $(d + p) \choose p$

- Gaussian: $K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$
  - every point is mapped to a Gaussian
France Telecom classifiers

Two detectors based on neural network

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

• Convolutional Face Finder (CFF) based on the use of a convolutional neural network
CGM Architecture

Decision

CGM 1
FRONT VIEW 1

CGM 2
FRONT VIEW 2

CGM 3
SIDE VIEW 1

CGM 4
SIDE VIEW 2

MLP
GATE

? Decision

grayscale image

MLP
FILTER

? Decision

color image

COLOR
FILTER

15 x 20 outputs

50 neurons

35 neurons

15 x 20 inputs

MMV
Multimedia and Vision Research Group

Queen Mary
University of London
INRIA’s approach

Training Images

Descriptors

Learning

Support Vector Machine

Test

Results

Multi-Scale detection

Test Image
Results

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)
Face Detection (QMUL)
Face Detection (FT)
Results (INRIA)
Features

QMUL
• High accuracy

FT/CFF is very fast
• Very efficient (330ms/image on a Pentium IV/3GHz)
• Low accuracy (sensitive against rotation)

INRIA
• Include additional features (body)
MOO of feature spaces

Improve discrimination power by combining multiple low-level features

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
Problems in feature combination

Advantage:
• Different feature spaces can be considered
• Features are weighted according to importance

Problems:
• Not all features are suitable for each concept
• ‘Averaging effect’
• Medium level performance

Distance combination
Feature Extraction

3 feature vectors
- Gabor textures
- Tamura textures
- Zernike moments

2 feature histograms
- SIFT-based Bag-of-features
- DCT-based Bag-of-features
Distance Measures

**Euclidean Distance**

$$d_2(x, y) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

**Histogram Intersection**

$$d_\cap(x, y) = \sum_{i=0}^{n} \min\{x_i, y_i\}$$

Different measures according to features structure.
Feature

Feature 1

Feature 2

Comparable
Feature Normalization

Normalize distance values to ensure a fair contribution of each visual feature

1. Identify the best PDF for distance values
2. Estimate parameters for that PDF
3. Normalize distances

Three step statistical normalization
Each feature follows its own distance distribution
PDF Estimation

Find the best fit using the KL-divergence between observed data and its estimated distribution.
Feature Combination Metric

Feature combination metric:

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

Weighting factors:

\{\omega_j, j \in F\}

How to find optimized weighting factors?
Training for concepts

Concept specific training

<table>
<thead>
<tr>
<th>$\nu_{1,1}$</th>
<th>$\nu_{1,2}$</th>
<th>....</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\nu_{2,1}$</td>
<td>$\nu_{2,2}$</td>
<td>....</td>
</tr>
<tr>
<td>$\nu_{3,1}$</td>
<td>$\nu_{3,2}$</td>
<td>....</td>
</tr>
<tr>
<td>....</td>
<td>....</td>
<td>....</td>
</tr>
</tbody>
</table>

Training set | Feature matrix | Multi-feature centroid
Multi-objective learning for weights

$v_{1,1} \quad v_{1,2} \quad d_{1,1} \quad d_{1,2} \quad \vdots \quad D_1 = \sum_{j=1}^{q} \omega_j d_{1,j}$
$v_{2,1} \quad v_{2,2} \quad d_{2,1} \quad d_{2,2} \quad \vdots \quad D_2 = \sum_{j=1}^{q} \omega_j d_{2,j}$
$v_{3,1} \quad v_{3,2} \quad d_{3,1} \quad d_{3,2} \quad \vdots \quad D_3 = \sum_{j=1}^{q} \omega_j d_{3,j}$

.... .... ..... .... .... ...

Objective functions

Feature matrix
Distance matrix

Training set
Multi-objective optimisation

Multiple objective functions

Conflicting interest

General optimum

Multiple objective optimisation

\[ D_1 = \sum_{j=1}^{q} \omega_j d_{1,j} \]
\[ D_2 = \sum_{j=1}^{q} \omega_j d_{2,j} \]
\[ D_3 = \sum_{j=1}^{q} \omega_j d_{3,j} \]
\[ \ldots \]

Objective functions

Retrieval in optimised multi-feature space

Optimised weighting factors

\[ \{\omega_j, j \in F\} \]

Optimised multi-feature metric

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

Query concept

Most similar images

Less similar images
Image Retrieval Performance

The graph illustrates the performance of different image retrieval methods, plotting average precision against recall. The methods compared are MOL, Linear, DCT, GT, SIFT, ZM, and TT. Each method is represented by a distinct line and marker, allowing for a visual comparison of their effectiveness.
Road Map

• Motivation
• Face detection and classifiers
• Face recognition
• Eigenfaces
• Social aspects (crowdsourcing)
• The human factor
• Latest from linear algebra
• Concluding remarks
Feature based matching

Input face (from face detector) → Normalization → Feature extraction

Feature vector

Output result
Name: Miss Maria Smith
Date of birth: 05.06.1978
Place of birth: London, UK
Etc …

Classifier/Decision maker

MMV
Multimedia and Vision Research Group
Queen Mary
University of London
Road Map

- Motivation
- Face detection and classifiers
- Face recognition
  Eigenfaces
  - Social aspects (crowdsourcing)
  - The human factor
- Latest from linear algebra
- Concluding remarks
The Space of Faces

- An image is a point in a high dimensional space
  - An N x M image is a point in $\mathbb{R}^{NM}$
Key Idea

- Images in the set $\chi = \{\hat{x}_R^L\}$ are highly correlated.

- Then, compress them to a low-dimensional subspace that captures key appearance characteristics.

- **EIGENFACES:** [Turk and Pentland]

USE PCA!
Eigenfaces look somewhat like generic faces.
Eigenfaces – summary

• Eigenfaces are the eigenvectors of the covariance matrix of the probability distribution of the vector space of human faces.

• Eigenfaces are the ‘standardized face ingredients’ derived from the statistical analysis of many pictures of human faces.

• A human face may be considered to be a combination of these standardized faces.
Generating Eigenfaces

1. Take a large set of images of human faces
2. The images are normalized to line up the eyes, mouths and other features.
3. The eigenvectors of the covariance matrix of the face image vectors are then extracted.
4. These eigenvectors are called eigenfaces.
Eigenfaces for Face Recognition

• When properly weighted, eigenfaces can be summed together to create an approximate gray-scale rendering of a human face.

• Remarkably few eigenvector terms are needed to give a fair likeness of most people's faces.

• Hence eigenfaces provide a means of applying data compression to faces for identification purposes.
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

\[
x \rightarrow ((x - \bar{x}) \cdot \mathbf{v}_1, (x - \bar{x}) \cdot \mathbf{v}_2, \ldots, (x - \bar{x}) \cdot \mathbf{v}_K) \\
\]

\[
a_1 \quad a_2 \quad \cdots \quad a_K
\]

\[
x \approx \bar{x} + a_1 \mathbf{v}_1 + a_2 \mathbf{v}_2 + \cdots + a_K \mathbf{v}_K
\]
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
    
    $$ x \rightarrow (a_1, a_2, \ldots, a_K) $$
  2. Given a new image (to be recognized) $x$, calculate $K$ coefficients
  3. Detect the closest labeled face in the database
Road Map

• Motivation
• Face detection and classifiers
• Face recognition
• Eigenfaces

Social aspects (crowdsourcing)
• The human factor
• Latest from linear algebra
• Concluding remarks
Joint People Recognition across Photo Collections using BN
Markus Brenner, Ebroul Izquierdo
Improve Face Recognition through Context

Face Recogn. → Context: e.g. Social Semantics

Extraction & Mining

Feedback
Detecting Body, Extracting Clothing
Exploiting Ambiguous Labels

Problem: known labels only given for sets of photos (faces), but not for particular faces

Approach: BN
Social Event Detection/Retrieval in Web Photo Collections

Social Events:
• Date and time
• Venue (geographic location)
• Involved people …
• … and their observable activities

→ Constraint-based spatio-temporal clustering and classification
Spatio-temporal propagation

- Not all photos include GPS data
- Our approaches is able to approximate the location if based on only textual information: by 22%
- Our approach is able to propagate and estimate the location if based on only the username: by 41%
- Our combined location propagation: by 64% to a total of 84%
Adding Crowdsourcing

1) Interactive web-based application: users can import photos (from local storage or internet)

2) Faces are detected and displayed (Dynamic (AJAX))
Gamification

2) Faces are detected and displayed.

3) Crowd-users play a game!

RESTful API
Demo available

Interested?
Road Map

- Motivation
- Face detection and classifiers
- Face recognition
- Eigenfaces
- Social aspects (crowdsourcing)
- The human factor
- Latest from linear algebra
- Concluding remarks
Latest work

• Background modeling:

  Given a set of images (from a video for instance), detect “invariant” parts (foreground)

• Image alignment problem + corruption detection:

  Given images of an object, align them to a fixed canonical templates and detect partial occlusions.
Model:

\[
\begin{bmatrix}
X \\
Data \ Matrix \ \ m \times n
\end{bmatrix} =
\begin{bmatrix}
L \\
Low \ Rank \ \ m \times n
\end{bmatrix} +
\begin{bmatrix}
S \\
Sparse \ Matrix \ \ m \times n
\end{bmatrix}
\]

\[
\begin{bmatrix}
I_1 & I_2 & \ldots & I_n
\end{bmatrix} =
\begin{bmatrix}
Stationary \\
Background \ \ m \times n
\end{bmatrix} +
\begin{bmatrix}
Moving \\
Objects \ \ m \times n
\end{bmatrix}
\]
Sequence Video

Stationary Background

Moving objects

Error: \( X - (L + S) \)
Calculating $L$ and $S$

- A first approach: Using Classical Principal Component Analysis

$L$ is the best rank-$r$ approximation of $X$:

$$L_r = \sum_{i=1}^{r} \lambda_i U_i V_i^T, \quad [U, S, V] = \text{svd}(X)$$

and the matrix $S$ is:

$$S = X - L$$

$L$ is a low-rank by construction but $S$ is not necessarily sparse.
Calculating L and S…

• A second approach: Consider the model \( X = L + S + G \)

Use a Lagrangian form:

\[
\min_{L,S} \left\{ \text{rank}(L) + \gamma \|S\|_0 \right\} \quad \text{st} \quad L + S = X
\]

\( \|S\|_0 \) : number of nonzero entries or cardinality of S.

This optimization problem is not tractable (nonconvexity of the objective function).
Calculating L and S…

- A third approach: (Candes, 2009, Peng, 2011)

  **Idea:** Impose a convex relaxation by changing the norms:

\[
\min_{L,S} \left\{ \|L\|_* + \gamma \|S\|_1 \right\} \quad \text{st} \quad L + S = X
\]

\[
\|S\|_* = \min_{i=1}^{\min(m,n)} \lambda_i(L) \quad , \quad \|S\|_1 = \sum_{i,j} S_{ij}
\]

The new objective function is non-smooth but now it is continuous and convex.

L and S are calculated using convex optimization
Calculating L and S…

• A fourth approach: (Zhou, 2011)

  Idea: Reduce the complexity fixing \( \text{rank}(L) \) and number of nonzero entries of \( S \).

\[
\min_{L,S} \left\{ \|X - L - S\|_F^2 \right\} \quad \text{st} \quad \text{rank}(L) \leq r, \text{card}(S) \leq k
\]

Main limitations:

• A restrictive condition: Cardinality should be fixed.

• High computing complexity: \( L \) and \( S \) are calculated by an iterative process where a SVD of the matrix \( X \) should be computed in each iteration.
Potential improvements:

Use randomized matrix algorithms for calculating a low rank approximation $L$.

What are the randomized matrix algorithms?

Techniques that use any randomized strategies to create a smaller problem where the standard methods can be applied.

Why is this so promising?

- Very fast algorithms.
- The use of the randomization has a regularizing effect with more robust outputs.
- They can exploit modern computational architectures better than classical numerical algorithms.
Road Map

• Motivation
• Face detection and classifiers
• Face recognition
• Eigenfaces
• Social aspects (crowdsourcing)
• The human factor
• Latest from linear algebra

Concluding remarks
Where are we now?

- False positives rates are close to zero for the best technologies and “good” conditions
- False negative rates are in the single digit range
- Detection failure rates are still higher than 12%
Where are we now?

Easy scenario:
Task can be solved with high accuracy
Where are we now?

Medium complexity - some R&D still needed
Where are we now?

High complexity: Crowds, many objects
– Substantial R&D needed–
The top 5 challenges (detection)

- Pose
- Illumination variations
- Expressions and occluding artifacts: glasses, huts, caps
- Clutter, crowds
- Picture quality: focus, distance, scale
Plus two more…

- Deliberated attacks (appearance change): Beard, hair cut, dark glasses, etc
- Privacy issues
- Deliberated physiognomy changes

See Kitanovski & Izquierdo 2011
Top 5 challenges (recognition)

Understand better the discriminative power of texture and shape?

Where are we now?

Yes Mr. Palmer, we here at Mega-Corp did try your "Face Recognition Software", and frankly, it did not work worth a dam...

Thanks