Face recognition in the wild: verification and caption-based recognition

Jakob Verbeek

LEAR Team, INRIA Grenoble, France

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

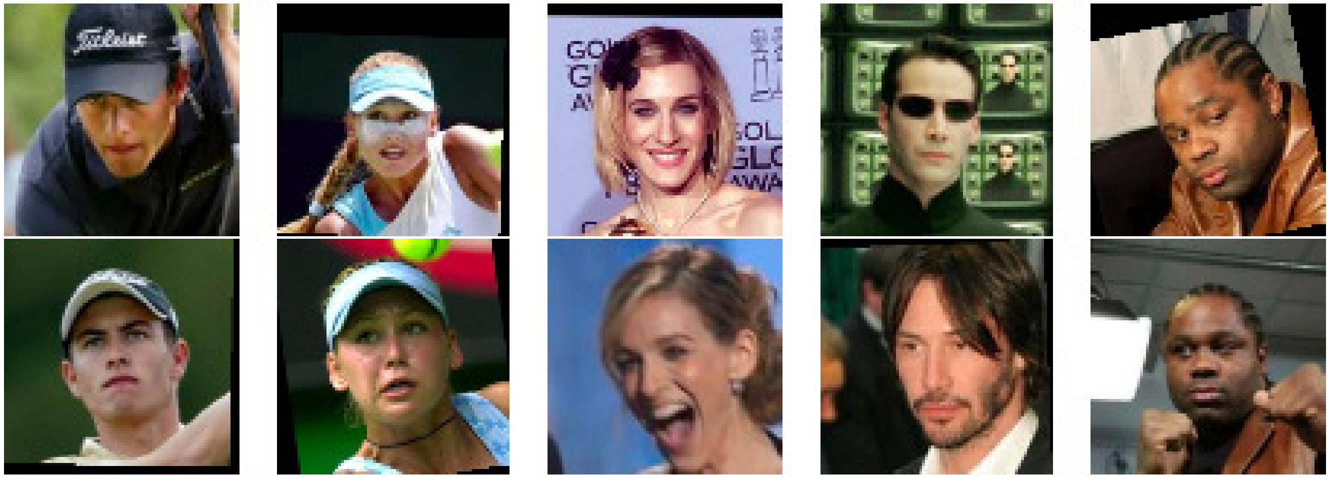
INRIA

Joint work with Matthieu Guillaumin (now at ETH Zürich) and Cordelia Schmid (LEAR)

From papers in: CVPR'08, ECCV'08, ICCV'09, ECCV'10, IJCV'11

Face verification

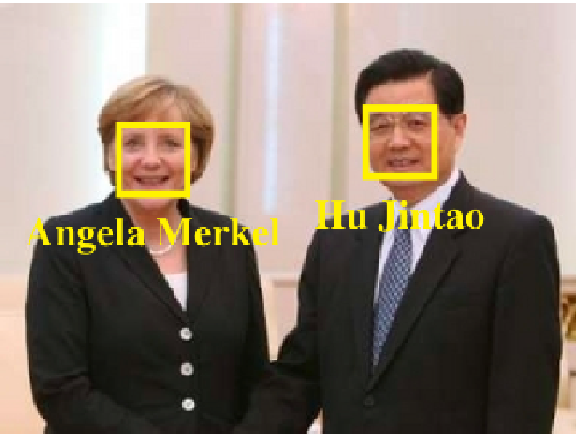
- Are these two faces of the same person?



- Challenges:
  - pose, scale, lighting, ...
  - expression, occlusion, hairstyle, ...
  - generalization to people not seen during training

Caption-based face recognition

- Identification without any labelled training examples **[Berg et al 2004]**
- Automatically detected faces from image, and names from caption



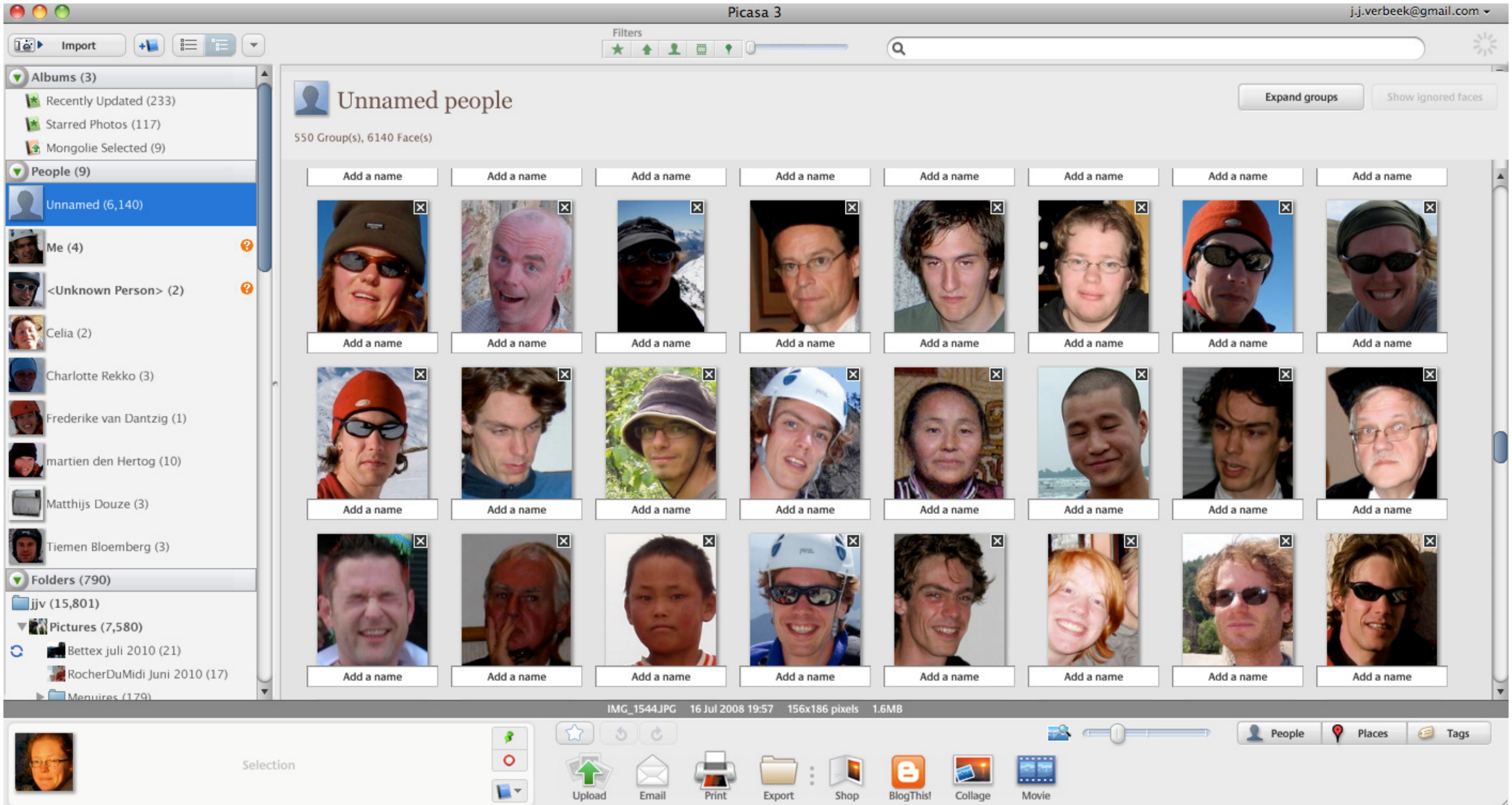
German Chancellor **Angela Merkel** shakes hands with Chinese President **Hu Jintao** (. . .)

Kate Hudson and **Naomi Watts**, Le Divorce, Venice Film Festival - 8/31/2003.

- Missed faces, erroneous face detections
- People not mentioned in caption, names missed

Unsupervised face clustering

- Example: grouping faces to speed-up labelling of personal photos

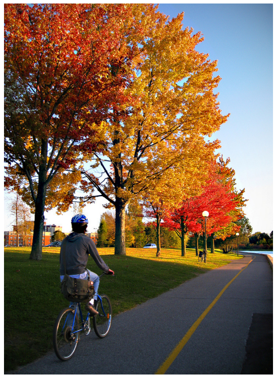


Picasa 3 screenshot

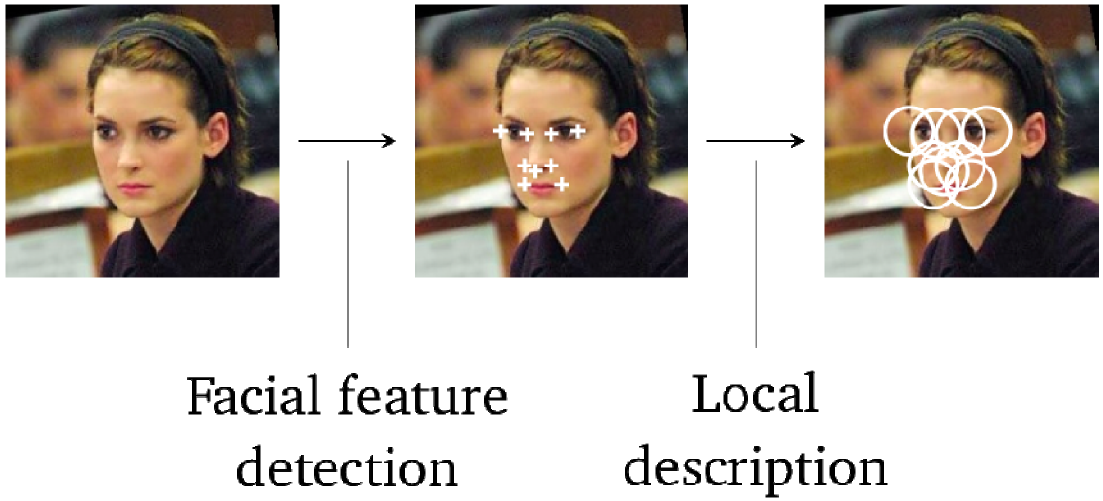
Metric Learning

- Acquisition of measures of distance or similarity from examples
- Which things are similar depends on the task at hand

season
scene type
objects

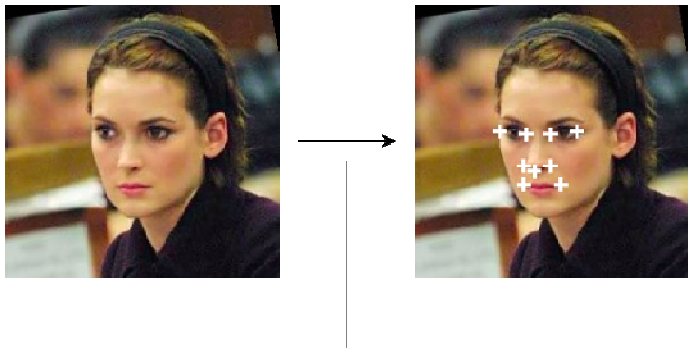


Feature extraction process

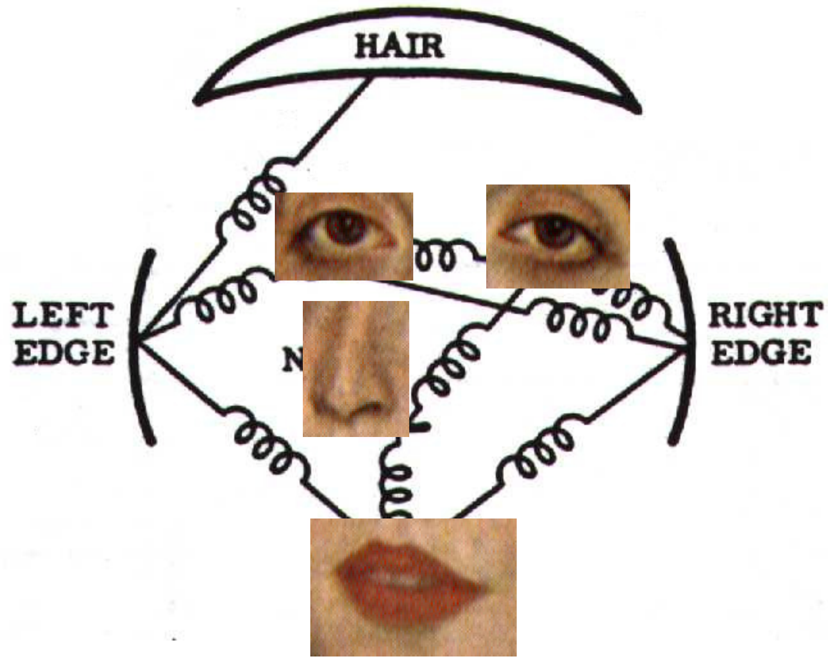


- Faces are not aligned, need features that are pose invariant
- Detection of 9 facial features using both appearance and relative position
- Each facial features described using SIFT descriptors

Feature extraction process (1)



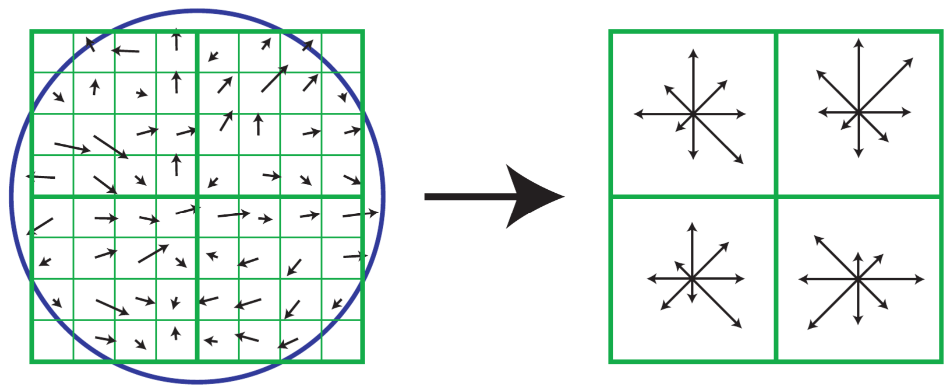
Facial feature detection



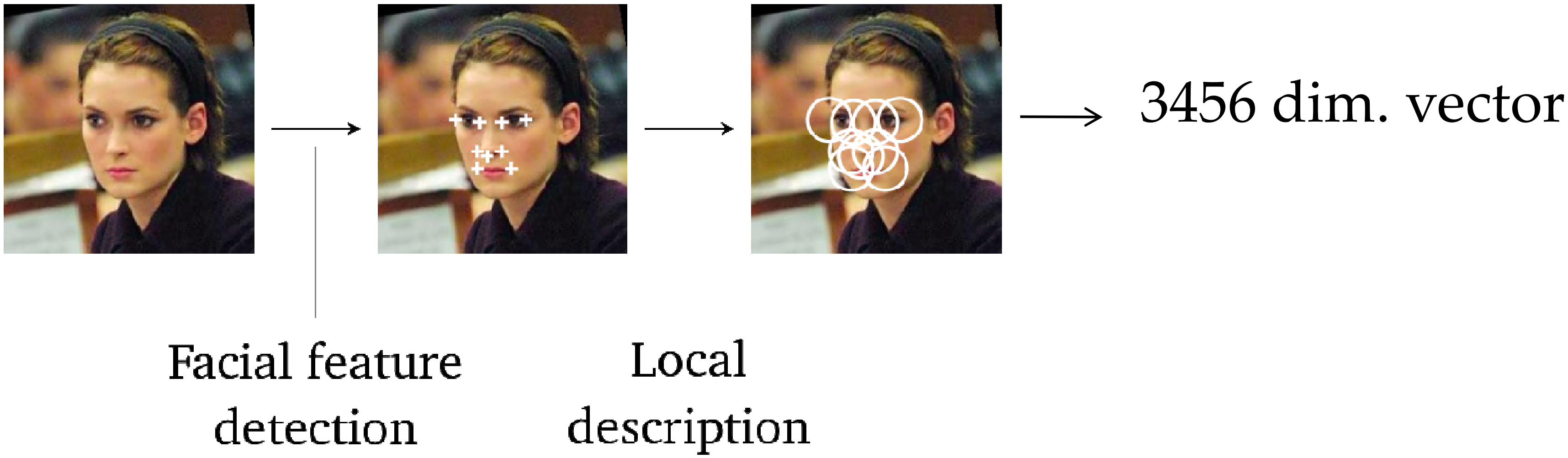
- Separate detectors for 9 facial parts: linear classifiers based on HoG features
  - learned from hand-annotated part locations
- Tree-structured model of quadratic displacement costs between parts
  - efficient identification of part locations using generalized distance transform

Feature extraction process (2)

- Each facial features described using 128d SIFT descriptors at 3 scales
- Concatenate 3x9 SIFTs into a vector of dimensionality 3456



Feature extraction process



- Detection of 9 facial features using both appearance and relative position
- Each facial features described using SIFT descriptors

- **Metric learning to find a distance useful for identification**

Metric Learning

- Euclidean or L2 distance is probably the most well known

$$d_{L2}(x,y) = (x-y)^T (x-y)$$

- Most common form of learned metrics are Mahalanobis

$$d_M(x,y) = (x-y)^T M (x-y)$$

  - M is a positive definite matrix
  - Generalization of Euclidean metric (setting M=I)
  - Corresponds to Euclidean metric after linear transformation of the data

$$d_M(x,y) = (x-y)^T M (x-y) = (x-y)^T L^T L (x-y) = d_{L2}(Lx, Ly)$$

- Clearly, not all methods fit this formulation of fixed vectorial data representation, eg based on matching image regions **[Nowak & Jurie 2007]**

Logistic Discriminant Metric Learning

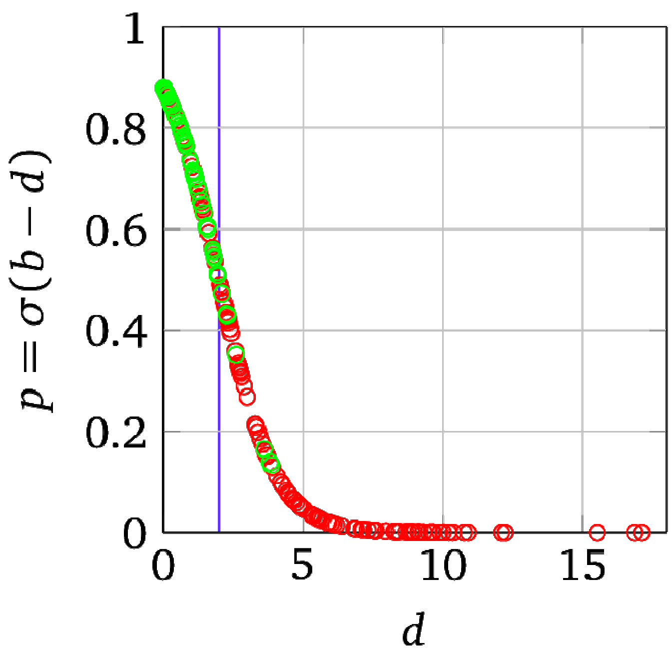
- Classify **pairs of faces** based on distance between descriptors

$$d_M(x_1, x_2) = (x_1 - x_2)^T M (x_1 - x_2)$$

- Use sigmoid to map distance to class probability **[Guillaumin et al, ICCV'09]**

$$p(y_{ij} = +1) = \sigma\big(b - d_M(x_i, x_j)\big)$$

$$\sigma(z) = \big(1 + \exp(-z)\big)^{-1}$$

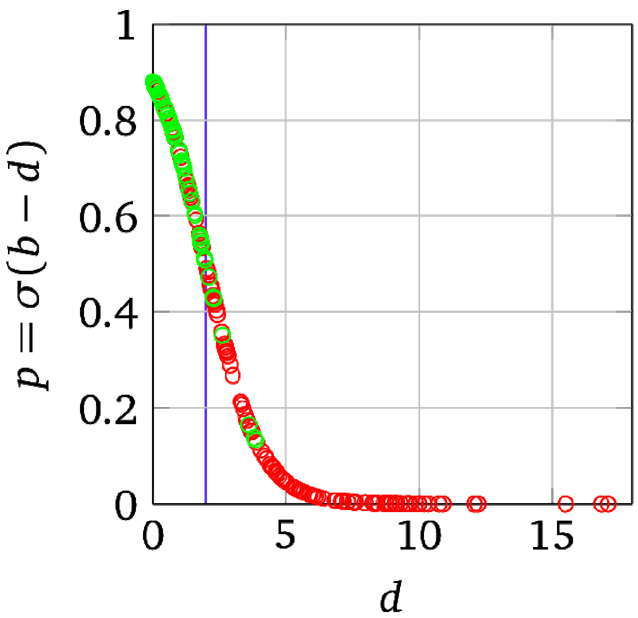


Logistic Discriminant Metric Learning

- Mahanalobis distance is **linear** in elements of M

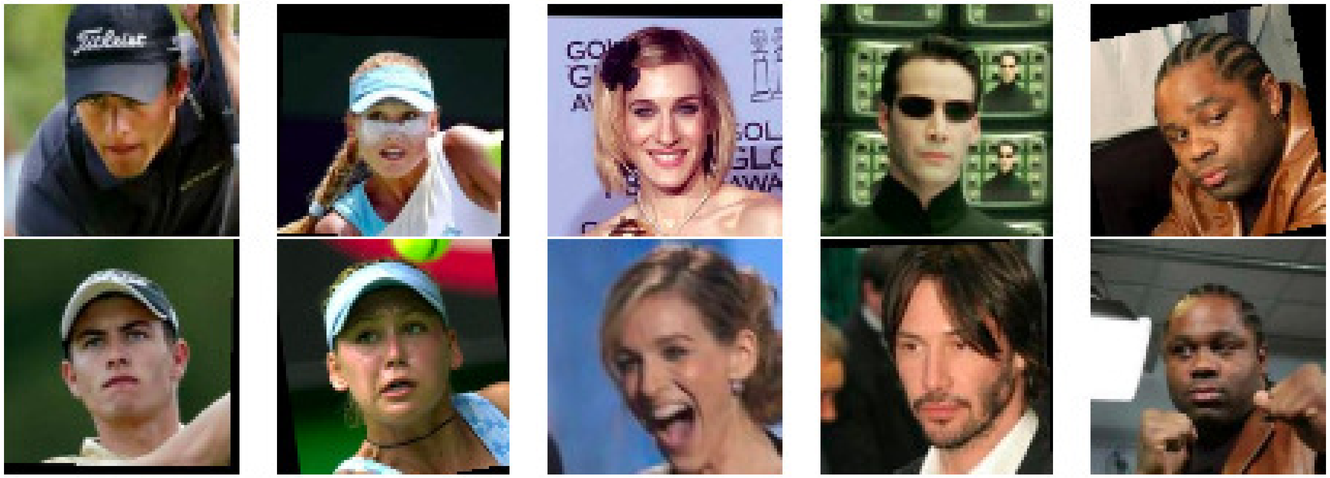
$$d_M(x_1, x_2) = (x_1 - x_2)^T M (x_1 - x_2)$$

$$p(y_{ij} = +1) = \sigma\left(b - d_M(x_i, x_j)\right)$$



- Standard **logistic discriminant** model
    - Learn maximum likelihood M and b
    - Convex optimization problem
- Can use **low-rank M =L^{T}L** to avoid overfitting
    - Loses convexity of cost function, but very effective in practice
    - Computational cost linear in dimension instead of quadratic !

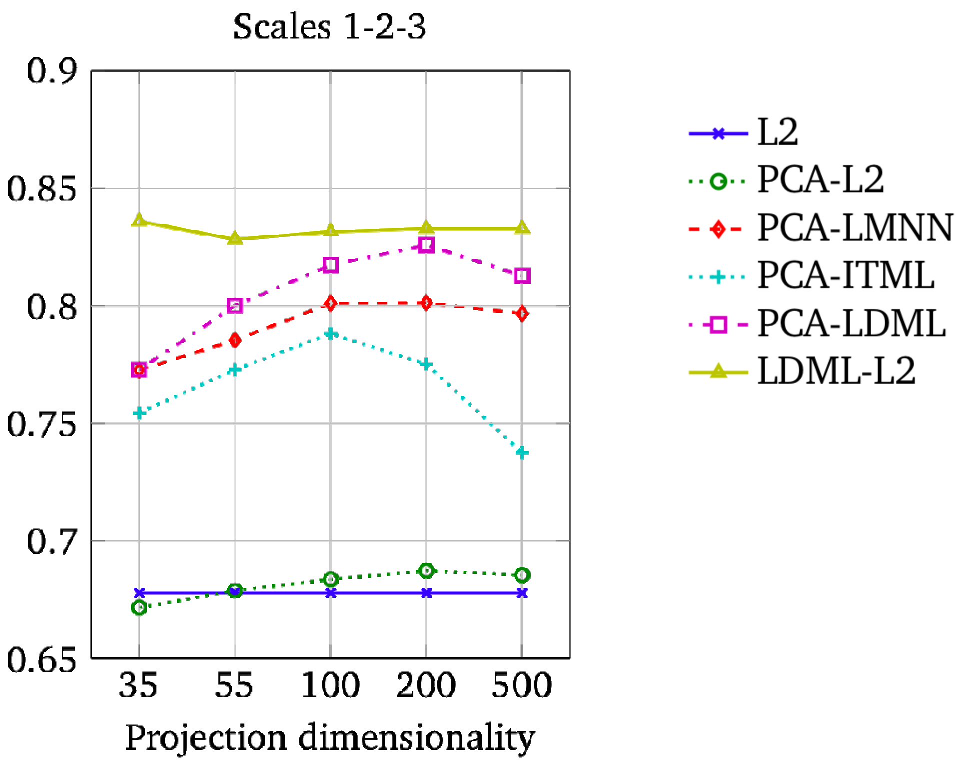
Labelled Faces in the Wild data set



- Contains 12.233 faces of 5749 different people (1680 appear twice or more)
- Realistic intra-person variability: pose, scale, lighting, expression, occlusion, …
- Detections from Viola & Jones detector, no proper alignment !
- People in test are not in the training set

Experimental Results

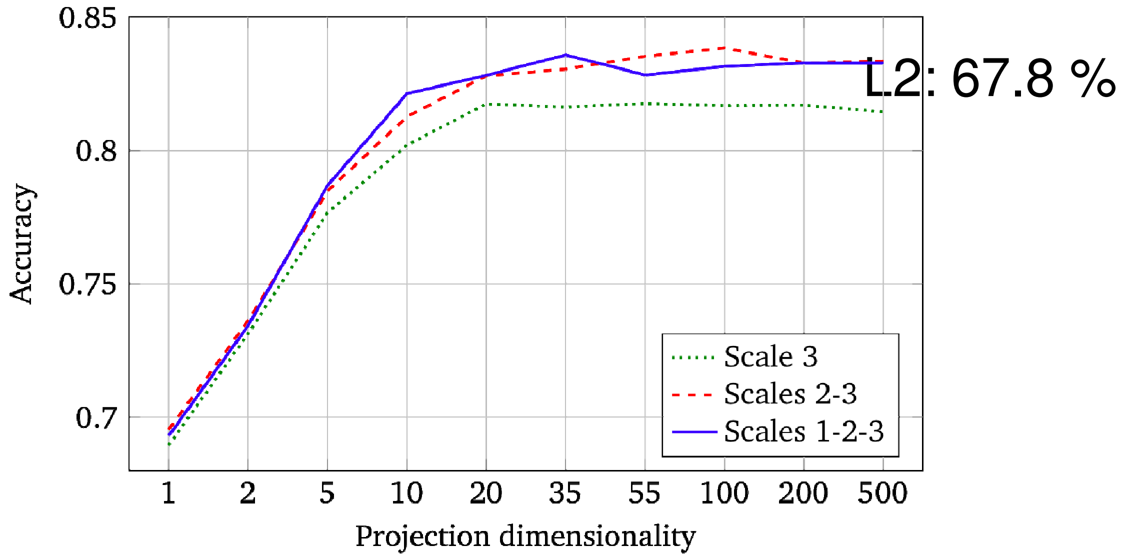
- Various metric learning algorithms on SIFT representation



  - Significant increases in performance when learning the metric
  - Learning low-rank metric better than chaining PCA and metric learning

Experimental Results

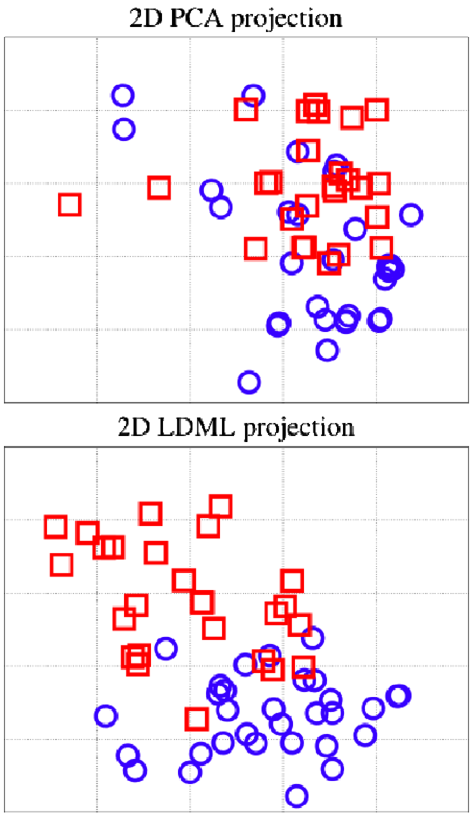
- Low-rank LDML metrics using various scales of SIFT descriptor



- Surprisingly good performance using few dimensions, like just 1 !
- Performance saturates already after around 20 of the 3456 dimensions

Comparing projections of LDML and PCA

• Using PCA and LDML to find two dimensional projection of the faces of Britney Spears and Jennifer Aniston



Marginalized k Nearest Neighbors

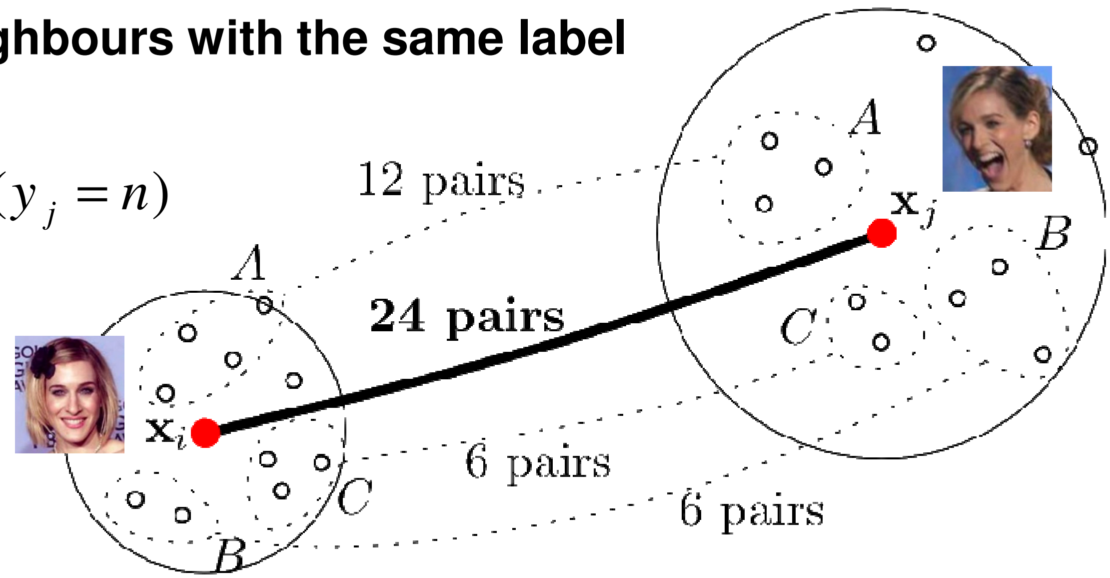
- Nearest neighbour prediction on identity of each face
  - Class probability given by fraction of neighbours of class

$$p(y_i = n) = c_{in} / k$$

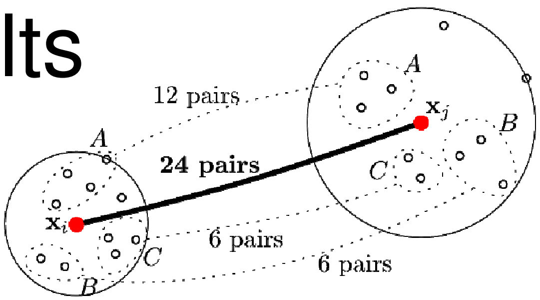
- Compute marginal probability that both samples belong to same class
  - **Counting pairs of neighbours with the same label**

$$p(y_i = y_j) = \sum_n p(y_i = n) p(y_j = n)$$

$$= \frac{1}{k^2} \sum_n c_{in} c_{jn}$$



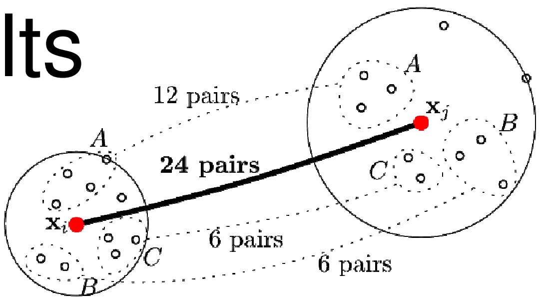
Marginalized kNN results



•Examples where LDML fails, but MkNN succeeds

• Cases with large variations in pose and expression

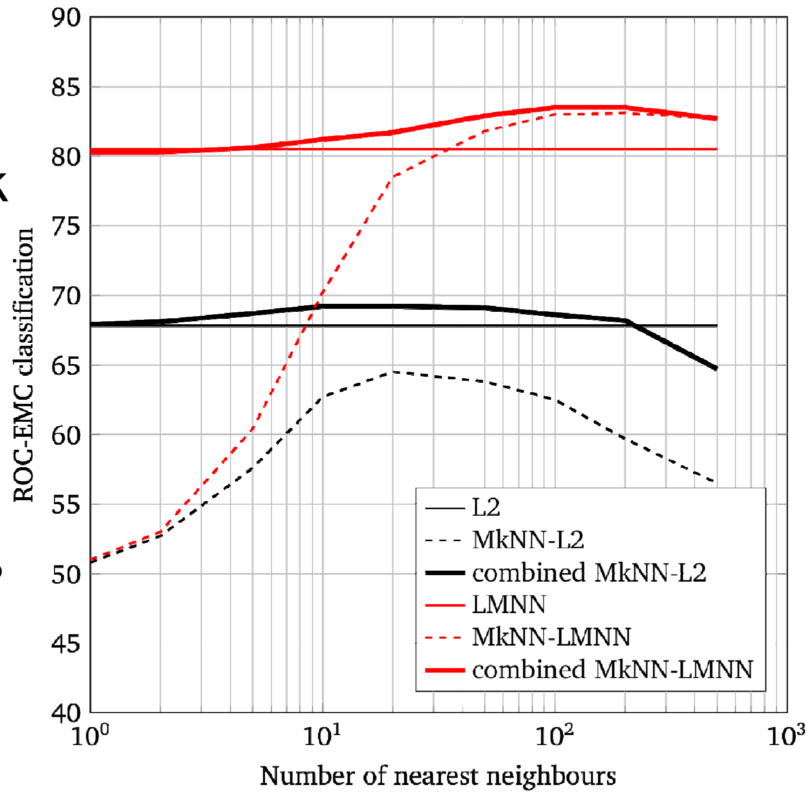
Marginalized kNN results



- Performance as function of
  - number of neighbours
  - Neighbour metric: L2 and LMNN

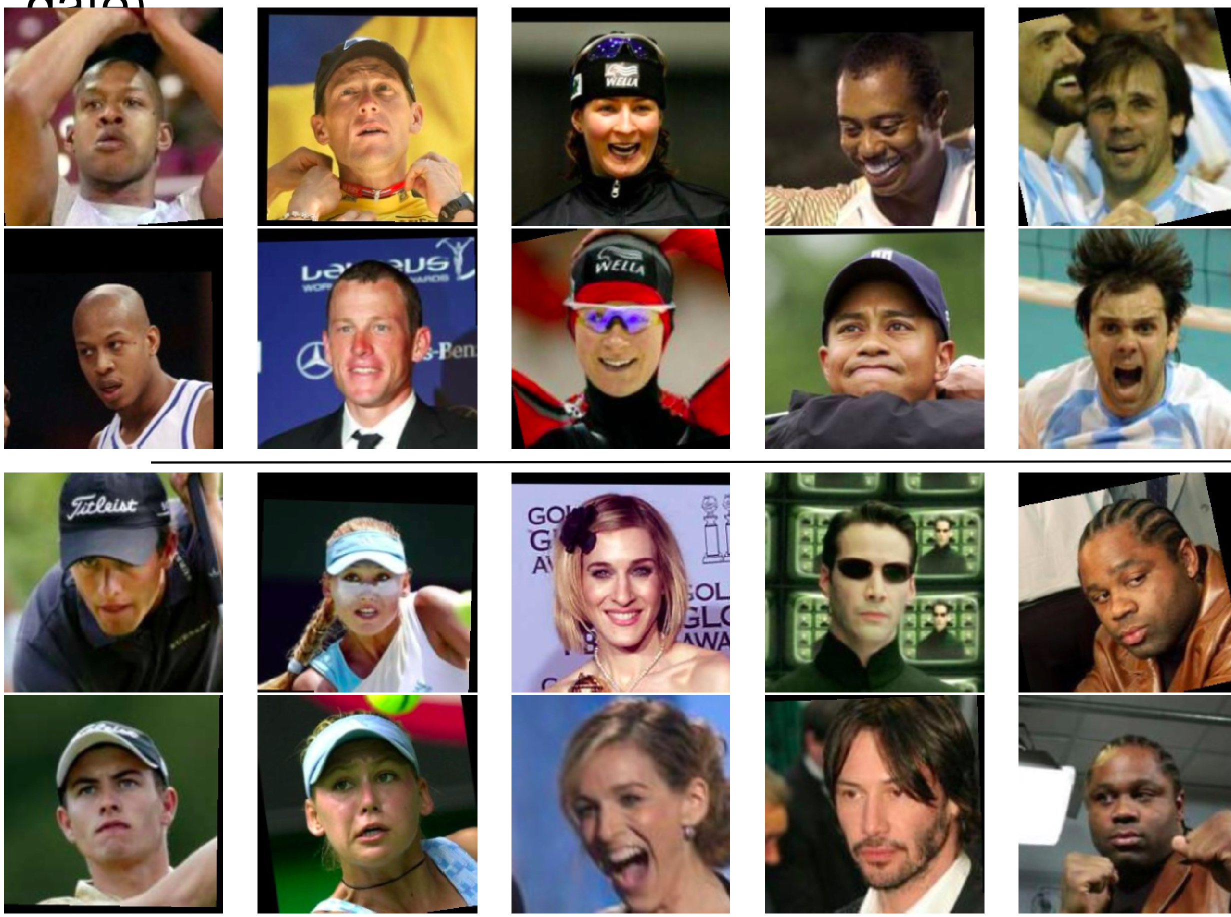
- Again: using the right metric for the task at hand is very important

- Performance comparable to LDML, methods complementary as a late fusion of the scores improves results to ~87.5%



Examples of face-pairs near decision boundary

- State of the art results on the LFW benchmark since 2009 (2^{nd} best to date)



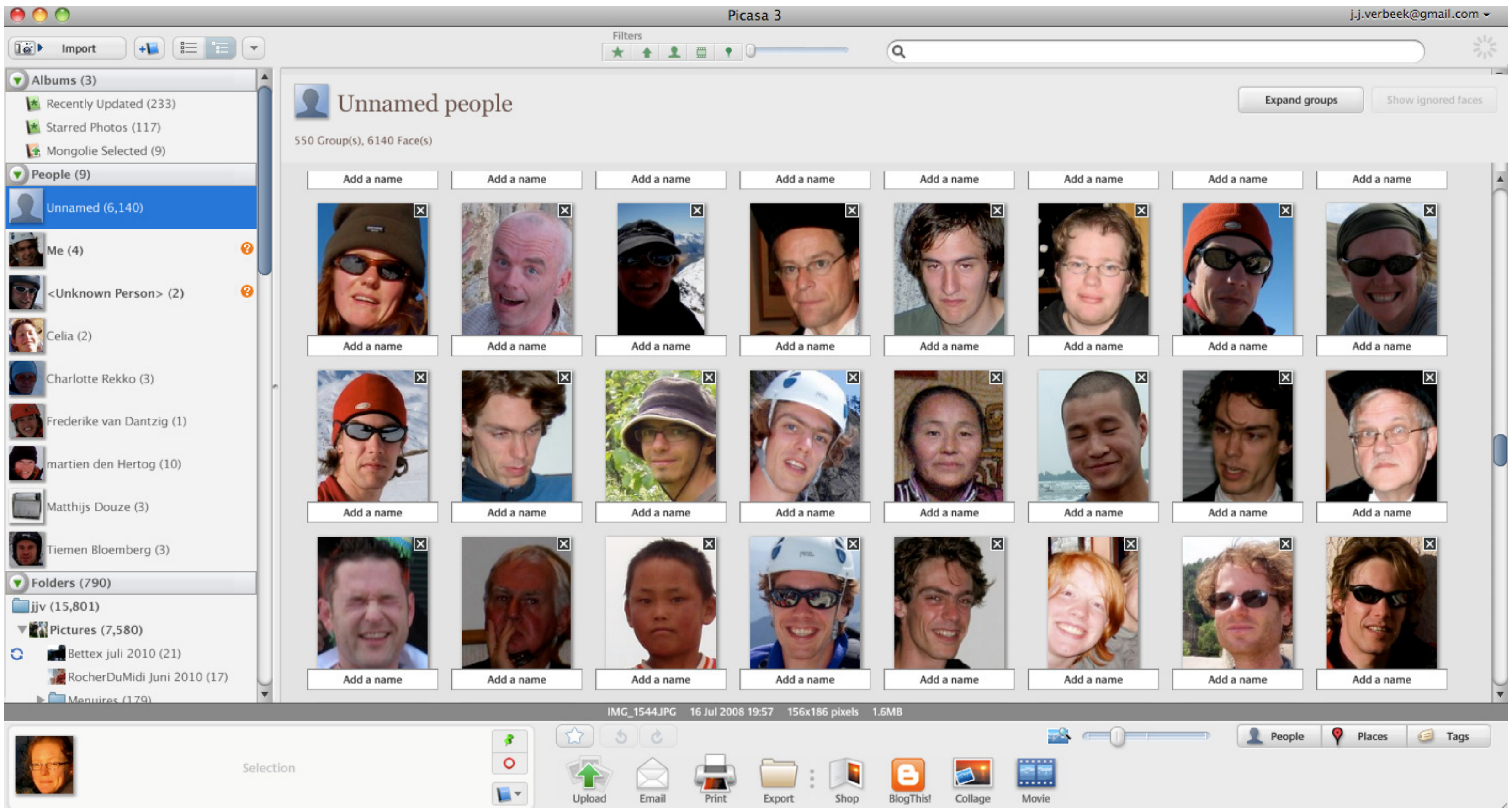
Correctly Classified

Incorrectly Classified

Application 1: Face Clustering

- Example: grouping faces to speed-up labelling of personal photos

Picasa 3 screenshot



Face Clustering experiment

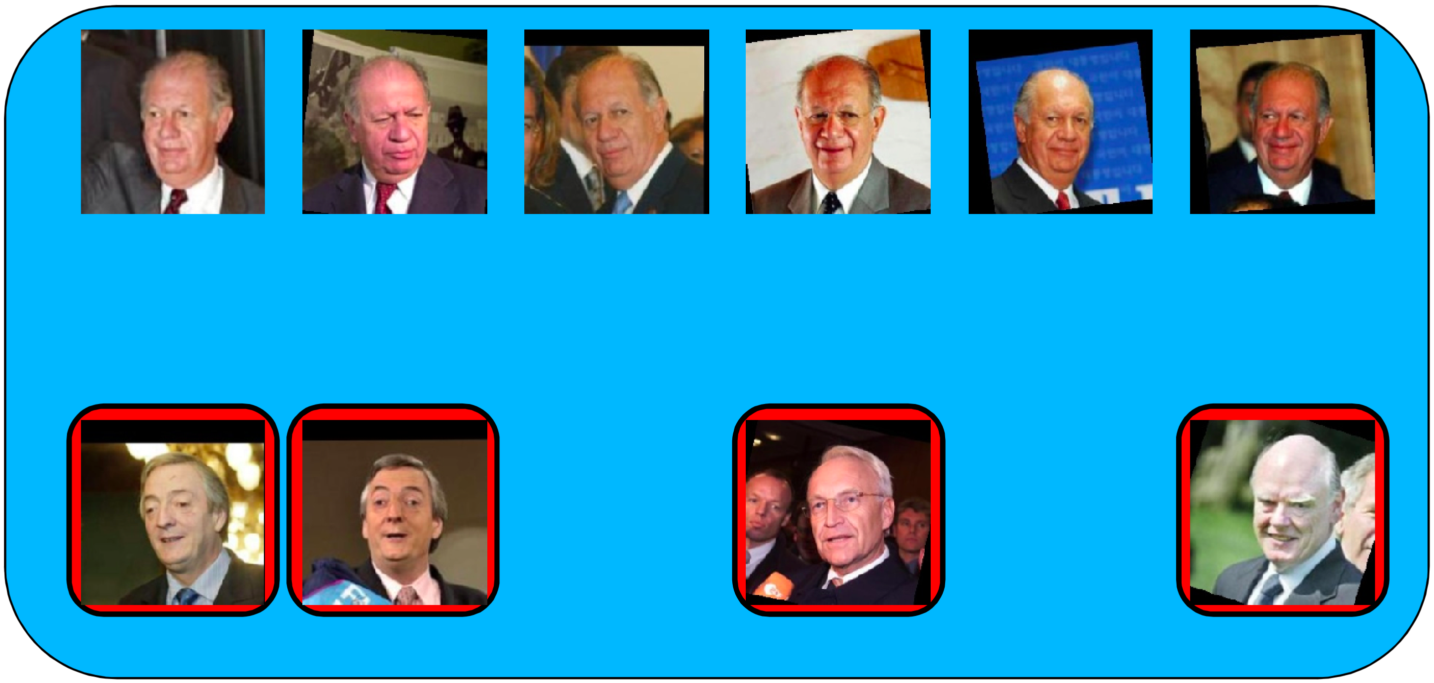
- Suppose user has two buttons
  - Button 1: Assign name to cluster of faces
  - Button 2: Assign name to a single face

- Labelling cost: number of clicks needed to name all faces

- Given a particular clustering, optimal labelling strategy
  - For each cluster
    - Assign cluster the name of most frequent person (1 click)
    - Correct all errors (1 click per remaining face)

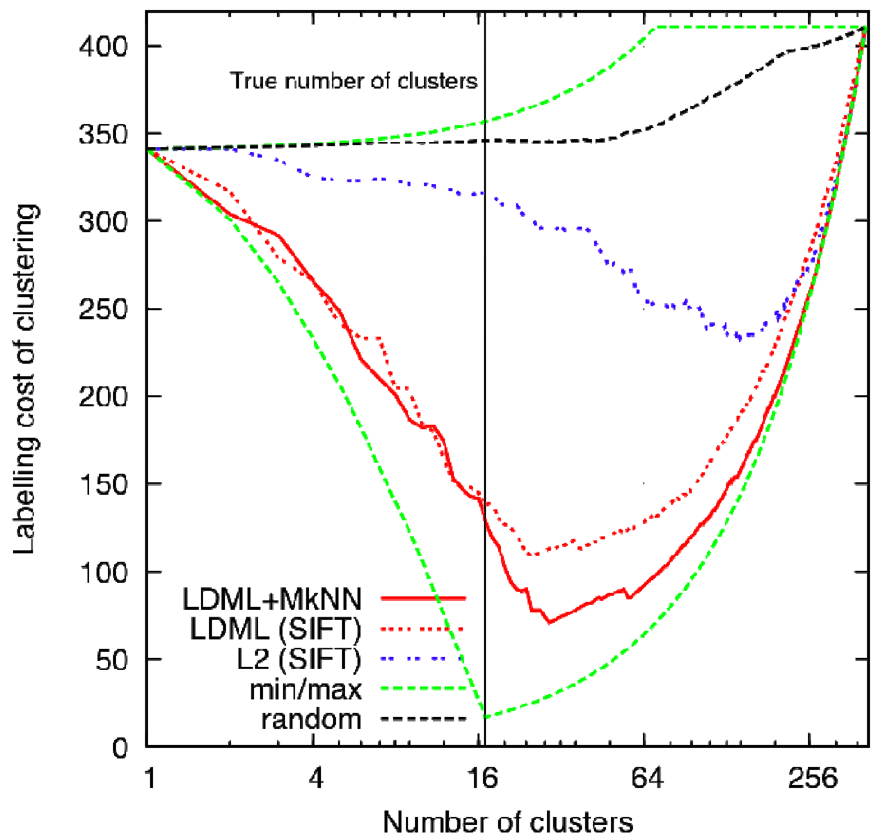
Face Clustering experiment

- Assign cluster the name of most frequent person (1 click)
- Correct all errors (4 clicks)



Face Clustering experiment

- Hierarchical clustering 411 faces of 17 people
- Varying the number of clusters



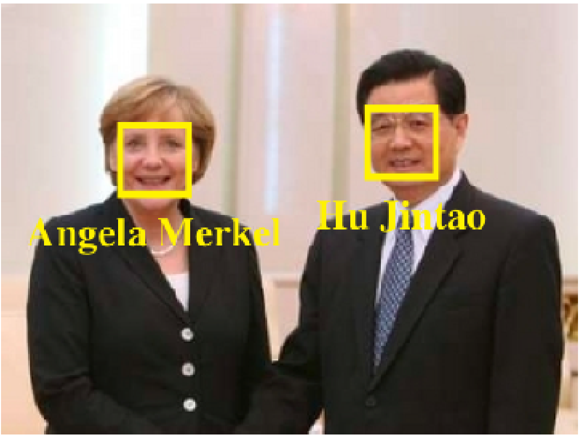
- Hierarchical clustering based on
  - L2
  - LDML (+MKNN)
  - random clustering
  - min/max labelling cost

- Learned metrics yield significantly better clustering results (6 faces per click vs 2 faces per click for L2)

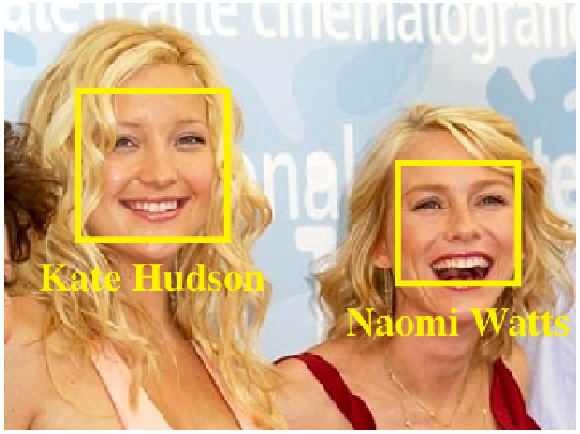
Example Clusters

Application 2: Caption-based recognition

- Identification without any labelled training examples **[Berg et al 2004]**
- Automatically detected faces from image, and names from caption



German Chancellor **Angela Merkel** shakes hands with Chinese President **Hu Jintao** (. . .)



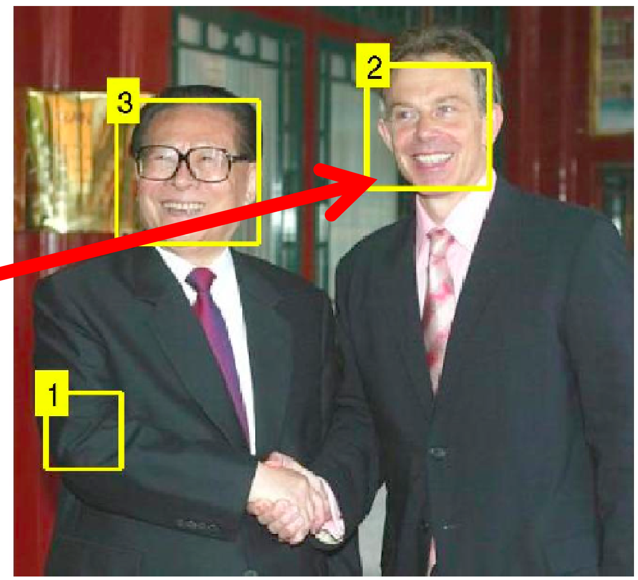
Kate Hudson and **Naomi Watts**, Le Divorce, Venice Film Festival - 8/31/2003.

- Missed faces, erroneous face detections
- People not mentioned in caption, names missed

Application 2: Caption-based recognition

- How can this work? By relying on a good face similarities !

George W. Bush
Tony Blair
Junichiro Koizumi



Tony Blair
David Kelly
Jiang Zemin

Labeled Yahoo! News data set

NEW!

- Subset of images gathered by Berg et al in 2002-2003 from Yahoo!News
- Kept 28.204 image with at least one detected face and name
- Manual annotation of each image indicates
  - Correct name-face associations
  - For unmatched name/face: is face missed, or not present.

- Train and test set, people never appear in both
  - Train: 10.709 images, 16.320 faces
  - Test: 9362 images, 14.827 faces

- Publicly available + face features, first to include image caption data

Caption-based face recognition

- Iteratively optimize name-face matching per image, keeping rest fixed
- Constraints on name-face assignments in an image-caption pair
    - People appear once per image
    - A face belongs to only one person
    - Faces only assigned to names in the caption, or discarded

Constrained Gaussian Mixture Model

- For each person in the database we model appearance with a Gaussian
- The discarded faces all modelled with a single "background" Gaussian

$$p(\{x_1,\ldots,x_F\}) = \sum_A p(A) \prod_{f=1}^{F} p(x_f \mid n) \qquad (n,f) \in A$$

- Faces in image modelled with MoG, constrained by set of assignments
    - Prior with single parameter to prefer "null" assignments
- Constrained Expectation-Maximization algorithm
    - E-step: find most likely admissible assignment of names to faces
    - M-step: update Gaussian models given new assignments
- Due to high dimensionality, covariance matrix constrained to diagonal

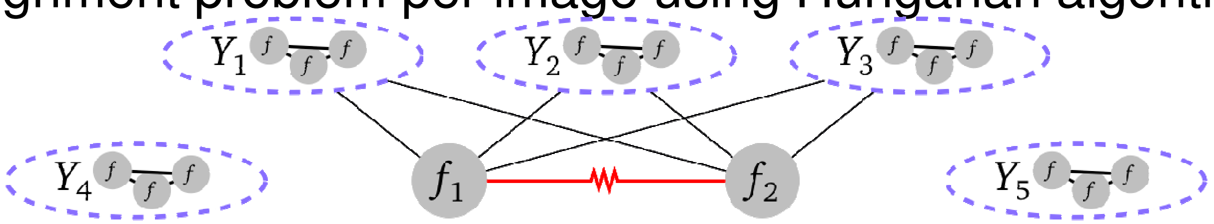
Direct similarity-based approach

- Maximize the sum of weights between faces assigned to same name
  - Weights given by –log of same-person probability estimate

- Fixed cost incurred by not assigning a face to a name **[Guillaumin et al. 2008]**

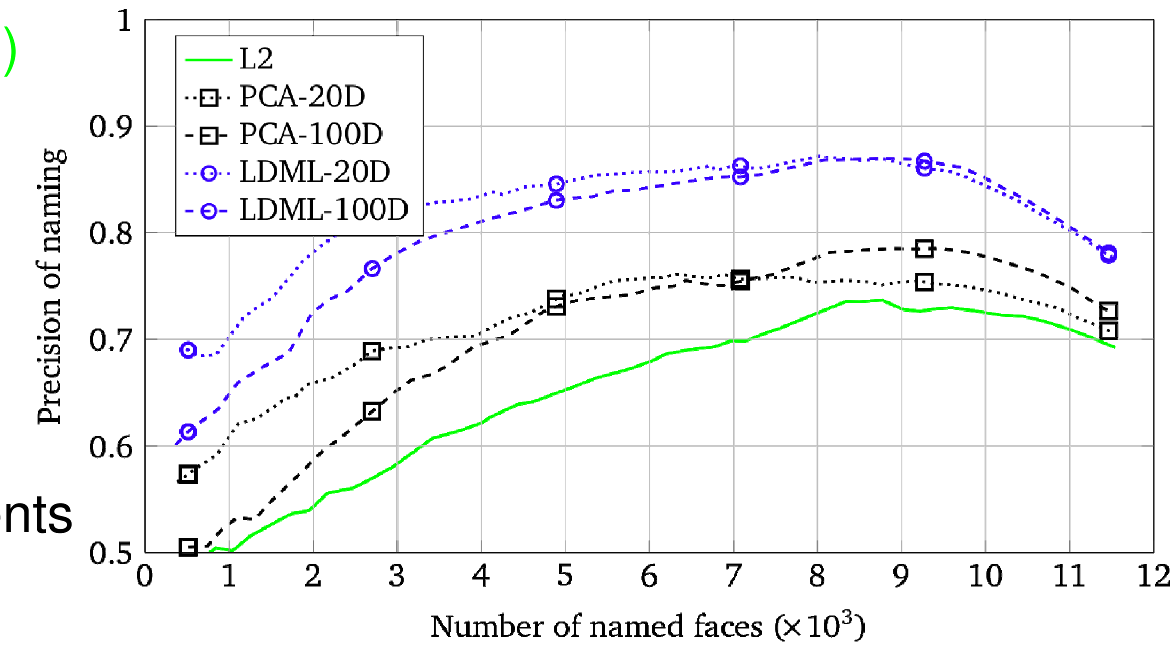
$$L(\{Y_n\}) = \sum_n \sum_{i \in Y_n} \sum_{j \in Y_n} w_{ij} + cN_\varnothing$$

- Compute for each face total sum of similarities for each possible name
- Solve assignment problem per image using Hungarian algorithm



Caption-based recognition experiments MoG

- Comparing mixtures learned in
  - Original space (L2)
  - PCA projection
  - LDML projection



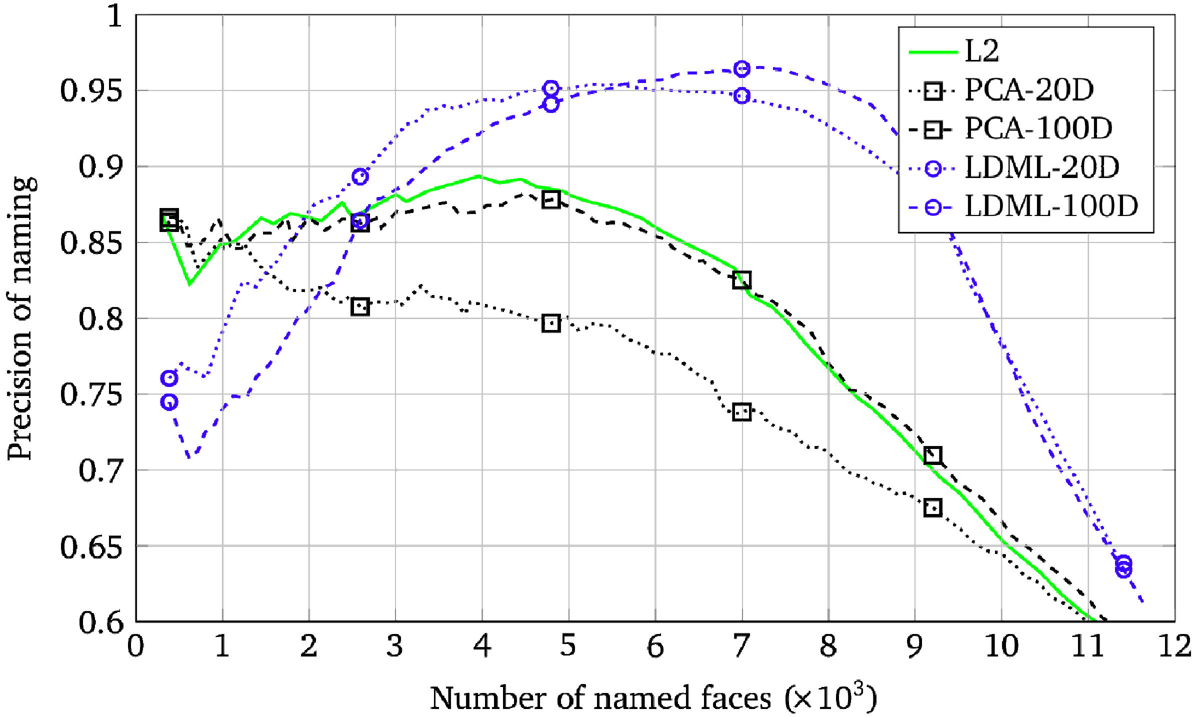
- Vary assignment prior

to prefer name assignments

- PCA helps: decorrelation
- LDML: suppresses irrelevant variations due to pose, expression, etc.

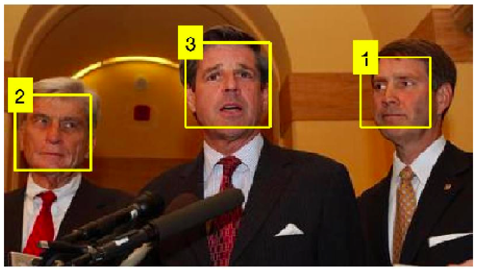
Caption-based recognition similarity-based

- Weights defined using distance from L2, PCA, LDML

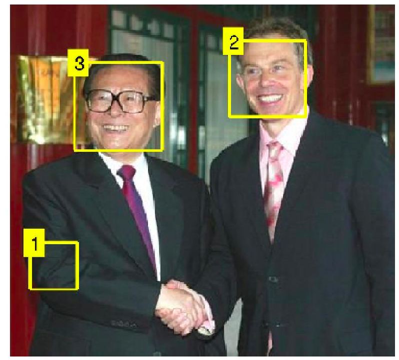


- PCA does not help: it preserves distances
- LDML: distances emphasise variations relevant for identity

Example name-face associations

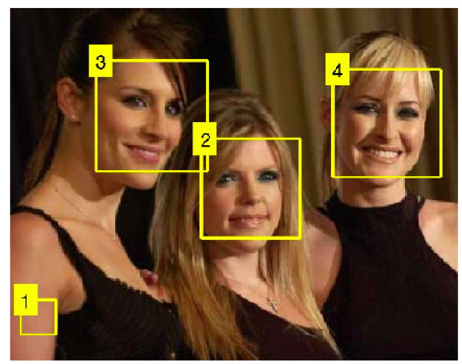


LDML	1. Saddam Hussein
	2. **John Warner**
	3. **Paul Bremer**
PCA	1. **Bill Frist**
	2. Paul Bremer
	3. Saddam Hussein



LDML	1. **null**
	2. **Tony Blair**
	3. **Jiang Zemin**
PCA	1. David Kelly
	2. **Tony Blair**
	3. **Jiang Zemin**

LDML	1. George W. Bush
	2. **null**
	3. **Tony Blair**
PCA	1. George W. Bush
	2. Junichiro Koizumi
	3. **Tony Blair**



LDML	1. **null**
	2. **Natalie Maines**
	3. **Emily Robison**
	4. **Martie Maguire**
PCA	1. **null**
	2. **Natalie Maines**
	3. Martie Maguire
	4. Emily Robison

Conclusions

- Metric learning significantly improvements performance of methods for verification, clustering, name-face associations

- Metric learning is also possible from weak caption-based annotation directly (see our ECCV'10 paper)

- Current challenges:
    - Dealing with occlusions of parts of the face
    - Matching faces under big pose changes: frontal vs. profile
    - Recognition, verification, clustering in video (TV series, movies)