Metric learning approaches for image annotation and face recognition

Jakob Verbeek

LEAR Team, INRIA Grenoble, France

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

INRIA

Joint work with :

Matthieu Guillaumin

Thomas Mensink

Cordelia Schmid

Similarities appear in many places in vision

- Matching: Distances between (local) image descriptors
  - wide baseline matching, image retrieval, …
- Clustering: Distance between data points and cluster centres
  - Visual dictionary construction, object discovery, …
- Classification: Kernels between images
  - Object recognition, localization, …

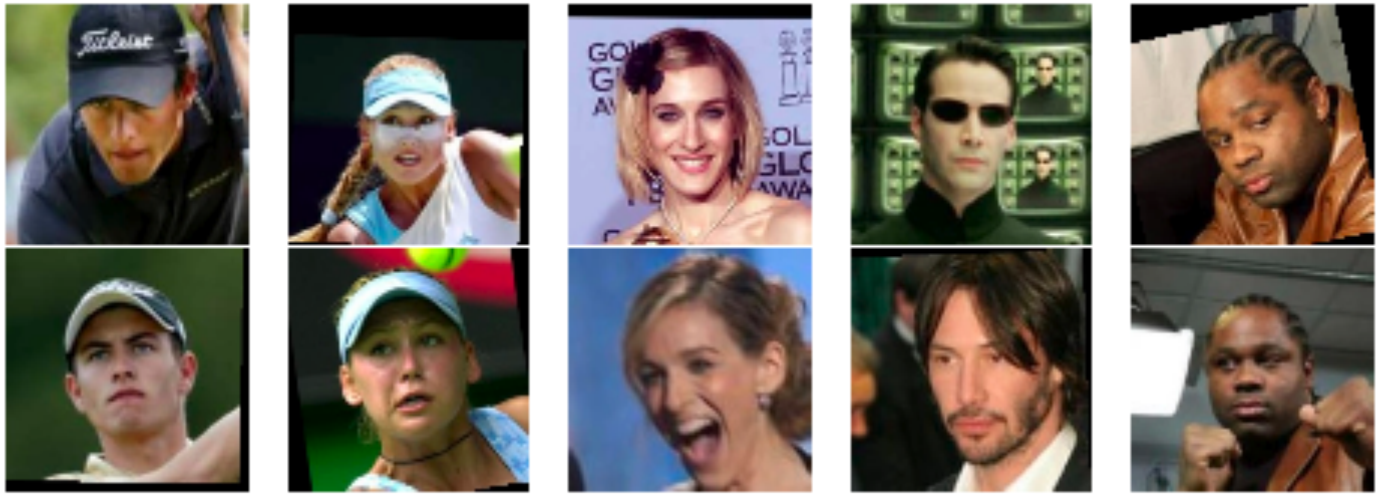
Metric Learning

- Acquisition of measures of distance or similarity from examples
- Which things are similar depends on the task
    - While visual features are often quite generic

season
scene type
objects

Learning metrics for face identification

- Are these two faces of the same person?



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

Metric learning for image annotation

- Predicting the relevance of keywords for images
  - Ranking keywords for an image
  - Ranking images for keywords

- Transfer annotations of visually similar training images

box	box (1.00)
brown	square (1.00)
square	brown (1.00)
white	white (0.79)
	yellow (0.72)

glacier	glacier (1.00)
mountain	mountain (1.00)
people	front (0.64)
tourist	sky (0.58)
	people (0.58)

blue	man (0.98)
cartoon	anime (0.96)
man	cartoon (0.92)
woman	people (0.89)
	woman (0.88)

landscape	llama (1.00)
lot	water (1.00)
meadow	landscape (1.00)
water	front (0.60)
	people (0.51)

Overview

1. **Metric learning methods**

2. Metric learning for image annotation

3. Metric learning for face identification
   - Application to face clustering
   - Application to caption-based recognition

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

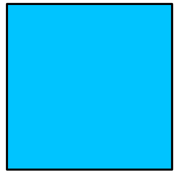
$$d_M(x,y) = (x-y)^T M(x-y) = (x-y)^T L^T L(x-y) = d_{L2}(Lx,Ly)$$

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

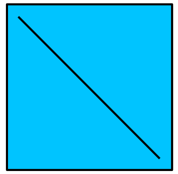
Metric Learning: different forms of M

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

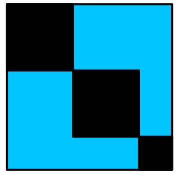
- Several popular choices for the form of M include
  - Full: quadratically many parameters



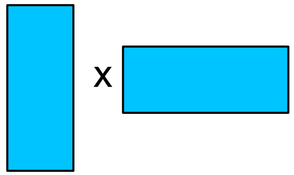
  - Diagonal: distance is a weighted sum of L2 distances computed on each dimension of the input vectors



  - Block diagonal: distance is a sum of Mahanalobis distances on different groups of dimensions (eg for different image descriptors)



  - Low rank: $M = L^T L$, where L is a (d x D) matrix, performs dimensionality reduction via linear projection



Metric Learning: different learning objectives

- Fisher's Linear Discriminant Analysis: linear projection to minimize within-class variance, maximize between-class variance **[Bishop 2006]**
  - Assumes Gaussian distribution of the data of each class

$$J(v) = \frac{v^T S_B v}{v^T S_W v}$$

$$S_W = \sum_k \sum_{n \in C_k} (x_n - m_k)^T (x_n - m_k)$$

$$S_B = \sum_k N_k (m - m_k)^T (m - m_k)$$

- Large Margin Nearest Neighbour metrics: force the nearest neighbours of each data point to be of the same class **[Weinberger et al 2005]**

$$E(M) = \sum_i \sum_{j \in N_i} d_M(x_i, x_j) + \sum_i \sum_{j \in N_i} \sum_{k \in R_i} \left[1 + d_M(x_i, x_j) - d_M(x_i, x_k)\right]_+$$

- Many more methods exist, for a recent survey see **[Yang & Jin 2006]**

Overview

1. Metric learning methods

2. **Metric learning for image annotation**

3. Metric learning for face identification
   - Application to face clustering
   - Application to caption-based recognition

Metric learning for image annotation

- Predicting the relevance of keywords for images **[Guillaumin et al 2009a]**
  - Ranking keywords for an image for (semi) automatic annotation
  - Ranking images for keywords to enable keyword based retrieval

- For test image transfer annotations of most similar training images

box	box (1.00)
brown	square (1.00)
square	brown (1.00)
white	white (0.79)
	yellow (0.72)

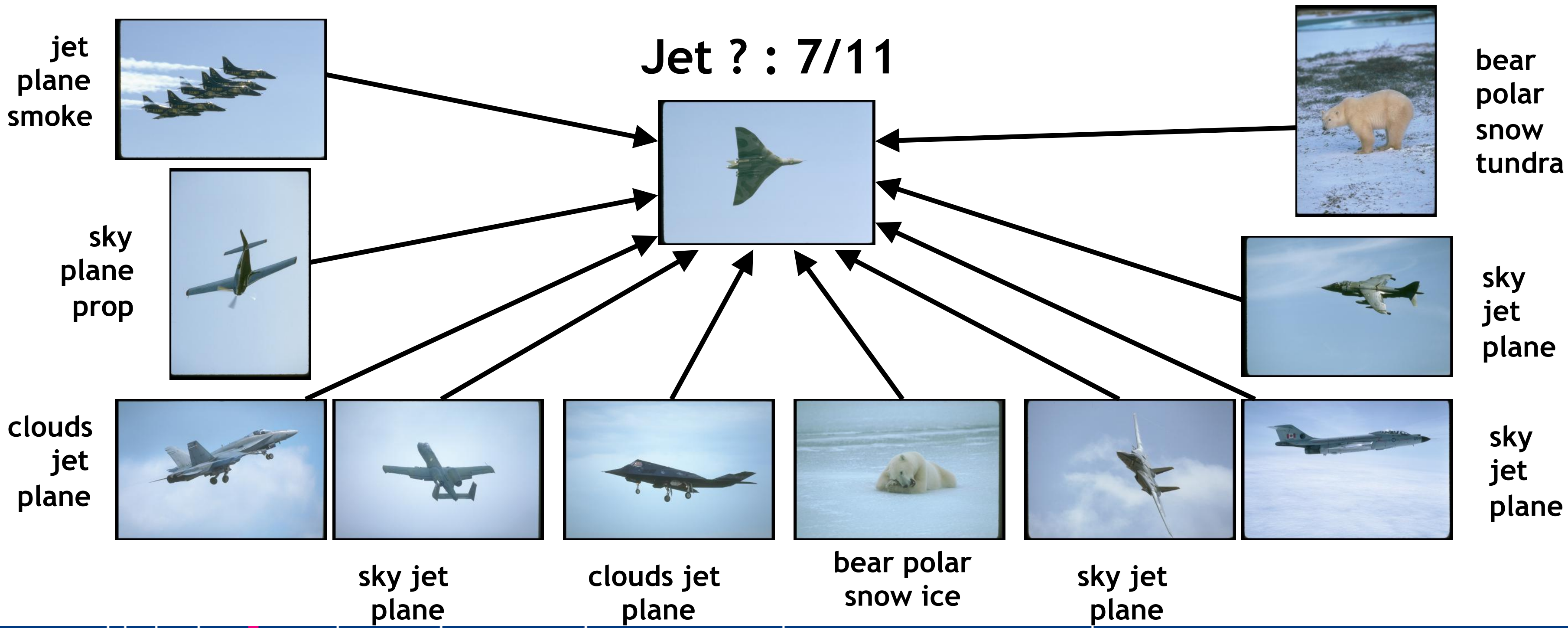
glacier	glacier (1.00)
mountain	mountain (1.00)
people	front (0.64)
tourist	sky (0.58)
	people (0.58)

blue	man (0.98)
cartoon	anime (0.96)
man	cartoon (0.92)
woman	people (0.89)
	woman (0.88)

landscape	llama (1.00)
lot	water (1.00)
meadow	landscape (1.00)
water	front (0.60)
	people (0.51)

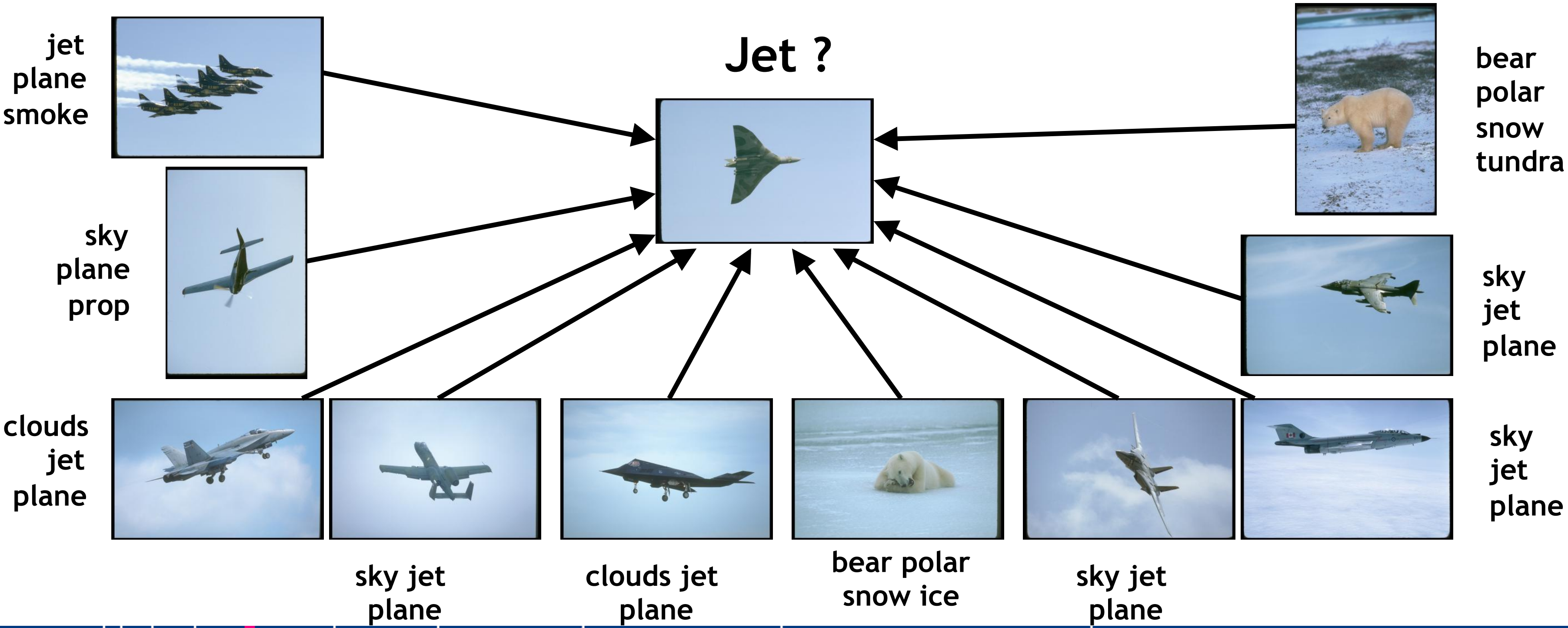
Nearest neighbour image annotation

- Take k neighbours, average their annotations
- Confidence score is fraction of neighbours annotated with the word



Nearest neighbour image annotation

- How many neighbours to use ?
- How to define neighbours, which distance ?



Nearest neighbour image annotation

- Nearest neighbour prediction for annotation bit $y \in \{0,1\}$

$$p(y=1) = \sum_{j=1}^{K} \frac{1}{K} y_{n_j}$$

Annotations of train images

- Generalizing Nearest Neighbour prediction
  - Relax the equal weighting of the k neighbours
  - Allow combination of multiple distances

$$p(y=1) = \sum_{j=1}^{N} \pi_j y_j$$

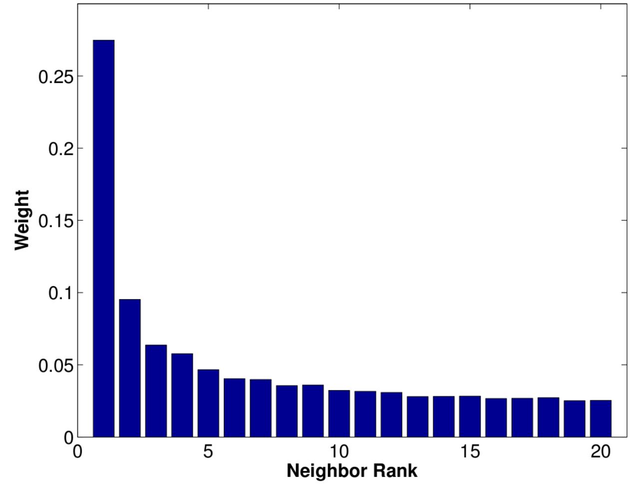
  - kNN obtained by setting weight to 1/K if j among K neighbours
  - Learn weights using maximum likelihood criterion

Rank-based nearest neighbour weights

- Prediction from weighted neighbours

$$p(y=1)=\sum_{j=1}^{N}\pi_j y_j$$

- Rank-based weighting: let r_j denote the neighbour rank of train image j
    - Learn weight w_k for each rank k



$$\pi_j = w_{r_j}$$

$$w_k \geq 0$$

$$\sum_k w_k = 1$$

- Multiple distances easily combined with weights for each combination of distance and rank

$$\pi_j = \sum_d w^d_{r^d_j}$$

$$w^d_k \geq 0$$

$$\sum_{k,d} w^d_k = 1$$

Distance-based nearest neighbour weights

- Prediction from weighted neighbours

$$p(y=1)=\sum_{j=1}^{N}\pi_j y_j$$

- Distance-based weighting: let d_j denote distances to train image j
  - Learn single parameter that sets decay of weight with distance

$$\pi_j=\frac{\exp(-wd_j)}{\sum_{j'}\exp(-wd_{j'})} \qquad w\geq 0$$

  - More generally we can learn a distance metric

$$\pi_j=\frac{\exp\left(-d_M(x,x_j)\right)}{\sum_{j'}\exp\left(-d_M(x,x_{j'})\right)}$$

Experimental evaluation

- Features extracted on each image, leading to image 15 distances
  - Gist descriptor
  - Colour histograms (3 color spaces, full image + 3x1 spatial grid)
  - Local descripors (SIFT + Hue, dense + Harris, full im + 3x1 grid)

- Metric learning: find weighted sum of 15 base distances

- Data sets

	Corel 5000	*ESP Game*	*IAPR TC-12*
Image size	256×384	variable	360×480
Vocabulary size	260	268	291
Number of training images	4500	18689	17665
Number of test images	499	2081	1962
Average number of words per img.	3.4	4.7	5.7
Maximum number of words per img.	5	15	23
Average number of img. per word	58.6	362.7	347.7
Maximum number of img. per word	1004	4553	4999

Image retrieval performance per keywords

- Performance of individual features, joint equal combination (JEC), and learned distance combination (MD), varying number of neighbours

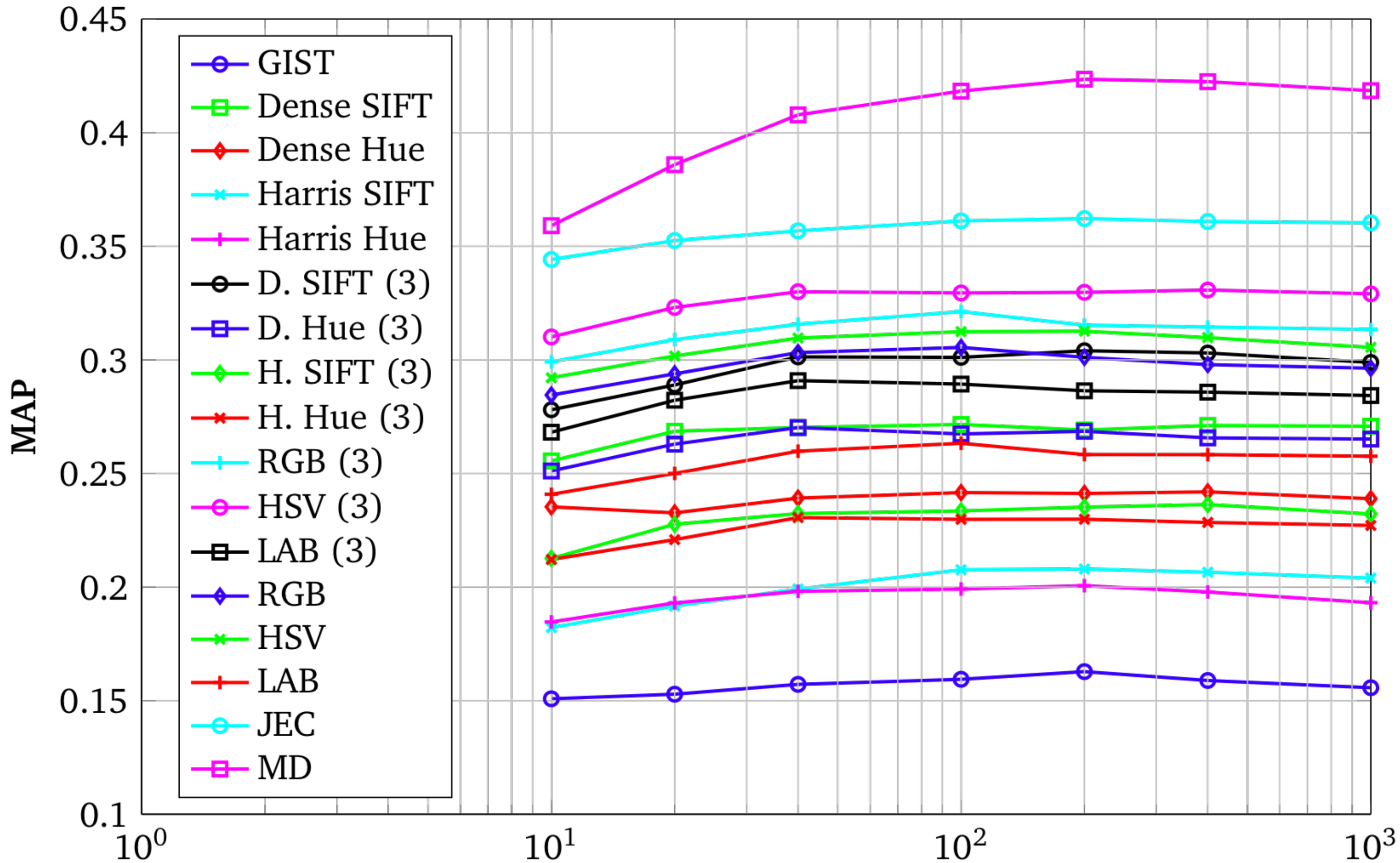
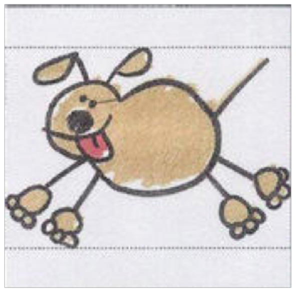


Image annotation examples

- Ground truth and predicted annotations (Correspondences in bold)

BEP: *40%*
Ground Truth: **wave** *(0.99)*, **girl** *(0.99)*, flower *(0.97)*, black *(0.93)*, america *(0.11)*
Predictions: people *(1.00)*, woman *(1.00)*, **wave** *(0.99)*, group *(0.99)*, **girl** *(0.99)*

BEP: *40%*
Ground Truth: **black** *(0.99)*, **picture** *(0.97)*, people *(0.97)*, painting *(0.90)*, group *(0.59)*
Predictions: old *(1.00)*, **black** *(0.99)*, gray *(0.99)*, man *(0.99)*, **picture** *(0.97)*



BEP: *40%*
Ground Truth: **drawing** *(1.00)*, **cartoon** *(1.00)*, kid *(0.75)*, dog *(0.72)*, brown *(0.54)*
Predictions: **drawing** *(1.00)*, **cartoon** *(1.00)*, child *(0.96)*, red *(0.94)*, white *(0.89)*

Overview

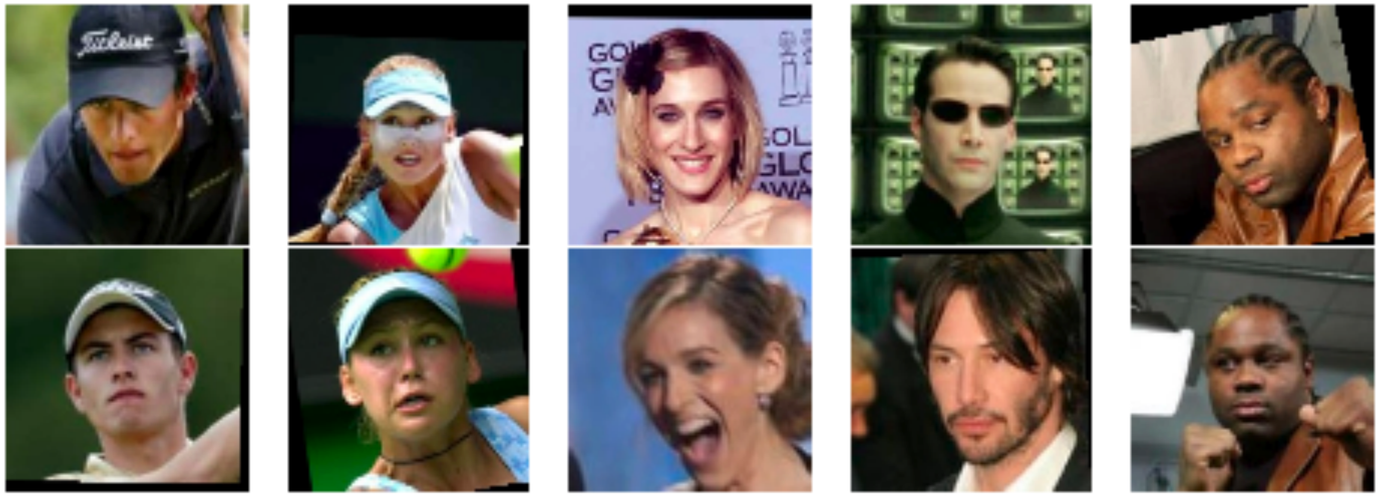
1. Metric learning methods

2. Metric learning for image annotation

3. **Metric learning for face identification**
    - Application to face clustering
    - Application to caption-based recognition

Learning metrics for face identification

- Are these two faces of the same person?

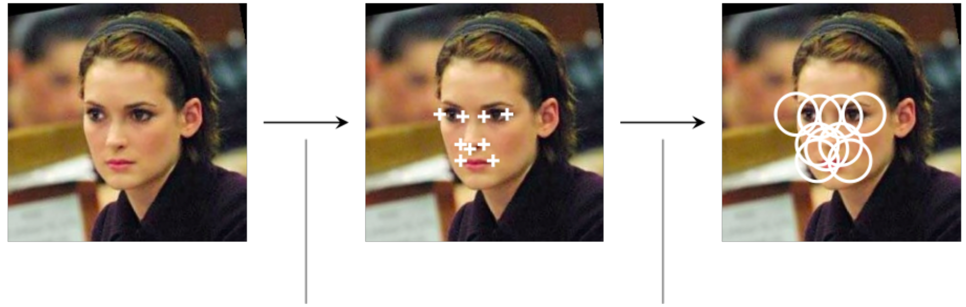


- Challenges:
  - pose, scale, lighting, ...
  - expression, occlusion, hairstyle, ...
  - generalization to people not seen during training

Face identification experiments

- Realistic intra-person variability: Labelled Faces in the Wild data set
  - Contains 12.233 faces of 5749 different people (1680 appear twice or more)
  - Task: predict for pair of faces whether they are the same person or not
  - Pairs used in test are of people not in the training set

- Feature extraction process



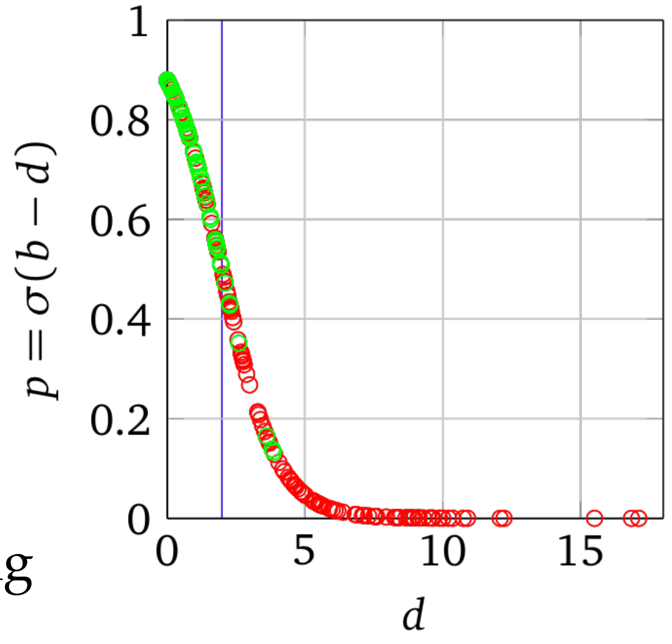
  - Detection of 9 facial features using both appearance and relative position **[Everingham et al. 2006]**
  - Each facial features described using SIFT descriptors at 3 scales
  - Concatenate 3x9 SIFTs into a vector of dimensionality 3456

Logistic Discriminant Metric Learning

- Classify **pairs of faces** based on a learned distance metric
- Use sigmoid to map distance to class probability **[Guillaumin et al 2009b]**

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

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

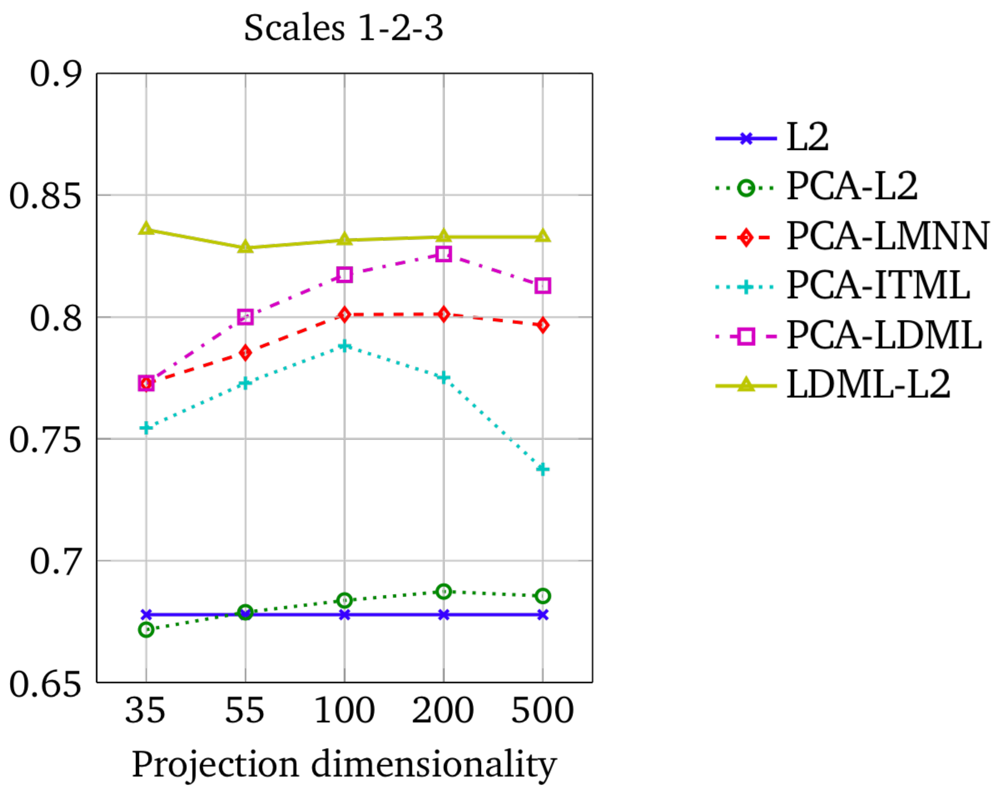


- Linear logistic discriminant model
  - Distance is linear in elements of M
  - Learn maximum likelihood M

- Can use low-rank $M = L^T L$ to avoid overfitting
  - Loses convexity of cost function

Experimental Results

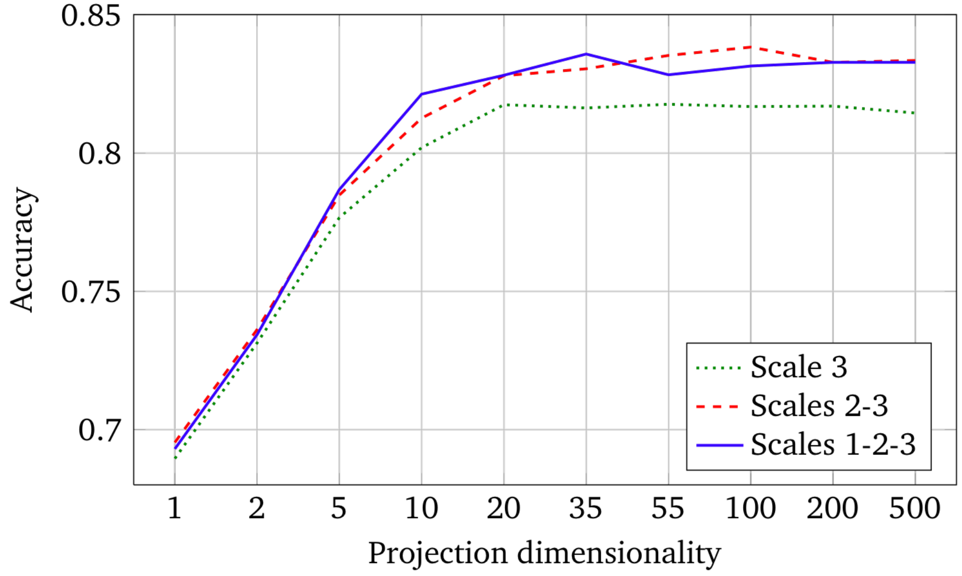
- Various metric learning algorithms on SIFT representation



  - Significant increases in performance when learning the metric
  - Low-rank metric needs less dimensions than PCA to learn good metric

Experimental Results

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



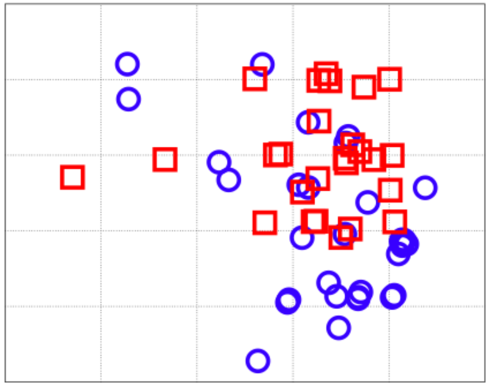
L2: 67.8 %

- Surprisingly good performance using very few dimensions
- 20 dimensional descriptor instead of 3456 dim. concatenated SIFT just from linear combinations of the SIFT histogram bins

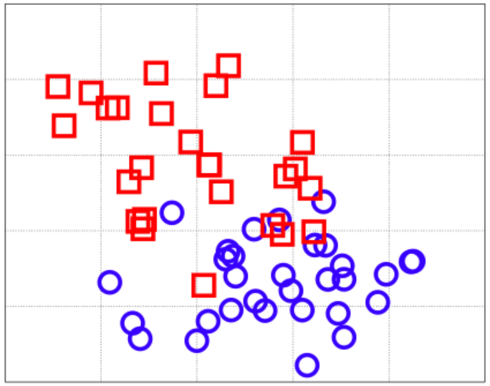
Comparing projections of LDML and PCA

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

2D PCA projection



2D LDML projection



Marginalized k Nearest Neighbors

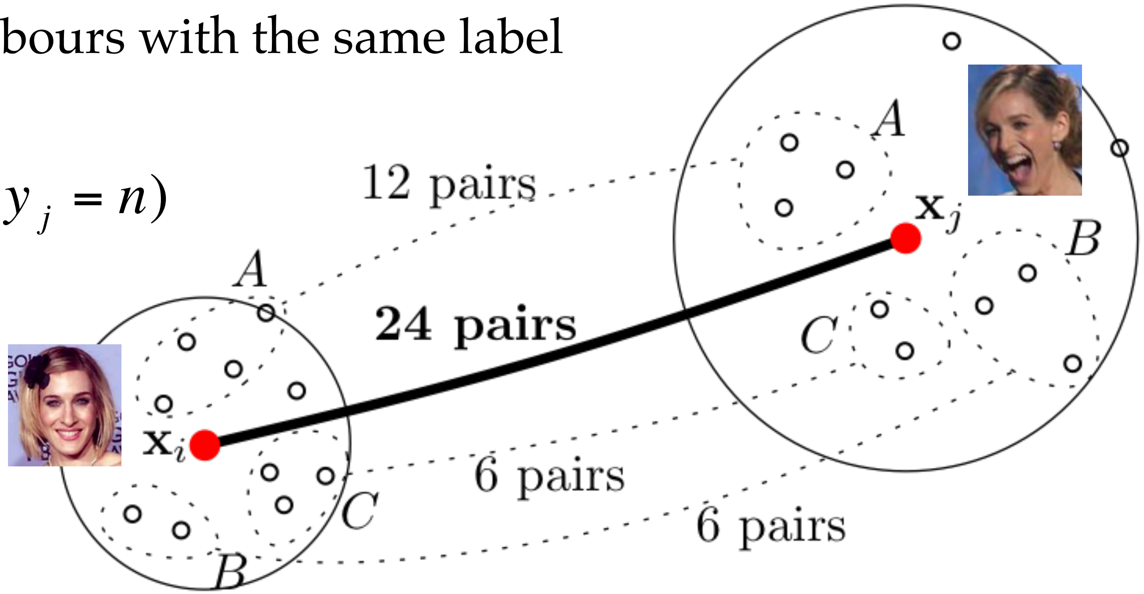
- Nearest neighbour prediction on identify each face
  - Class probability given by fraction of neighbours of class

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

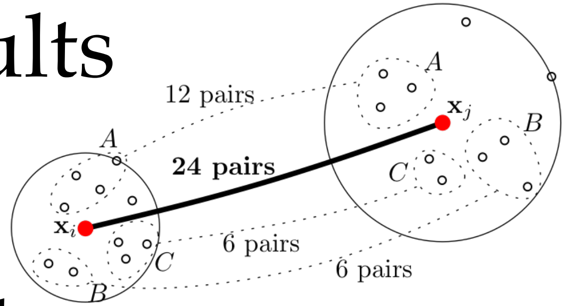
- Compute marginal probability that both samples belong to same class
  - Counting pairs of neighbours with the same label

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

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



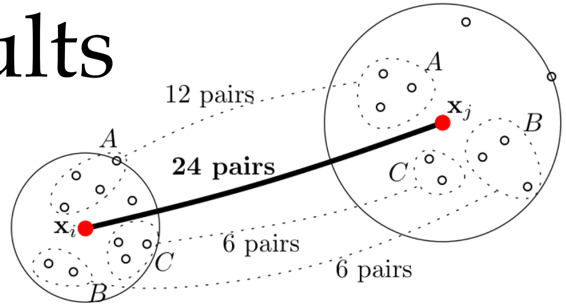
Marginalized kNN results



- Examples where LDML fails, but MkNN succeeds

- Observe the large variations in pose, expression

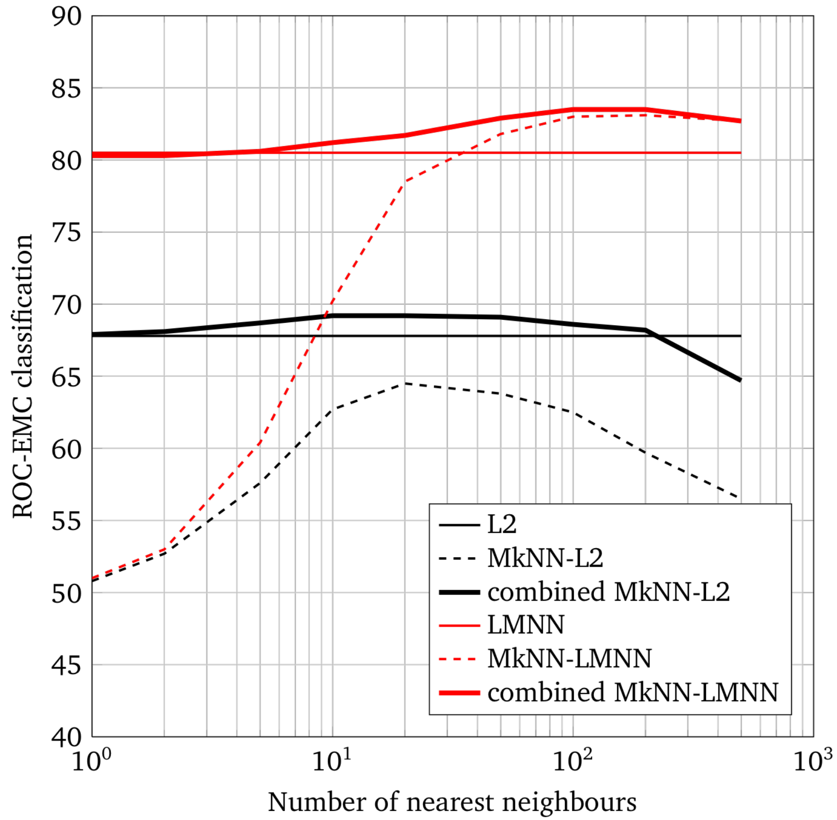
Marginalized kNN results



- Performance as function of
  - number of neighbours
  - Neighbour metric L2 / LMNN

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

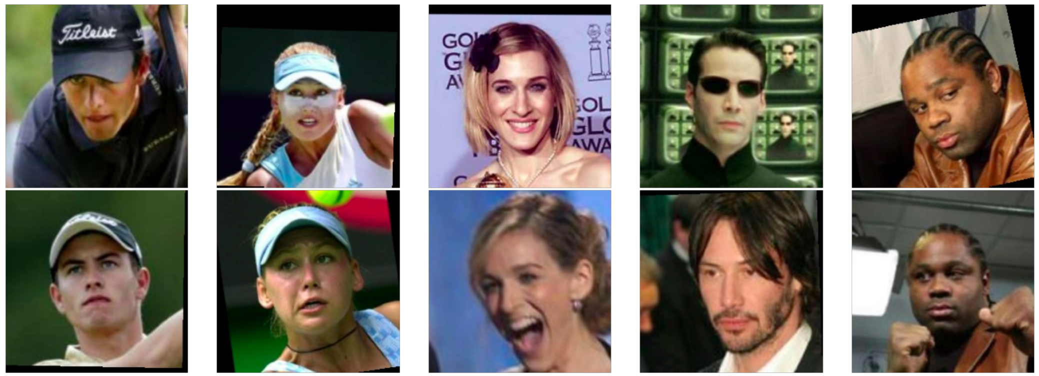
- Performance comparable to LDML, methods complementary as a late fusion of the scores improves results to ~87.5%



Examples of face-pairs need decision boundary

- Combining scores of LDML and MkNN further increases performance
  - State of the art results on the LFW benchmark

Correctly Classified



Incorrectly Classified

Overview

1. Metric learning methods
2. Metric learning for image annotation
3. Metric learning for face identification
    - **Application to face clustering**
    - Application to caption-based recognition

Application 1: Face Clustering

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

Picasa 3 screenshot

Face Clustering experiment

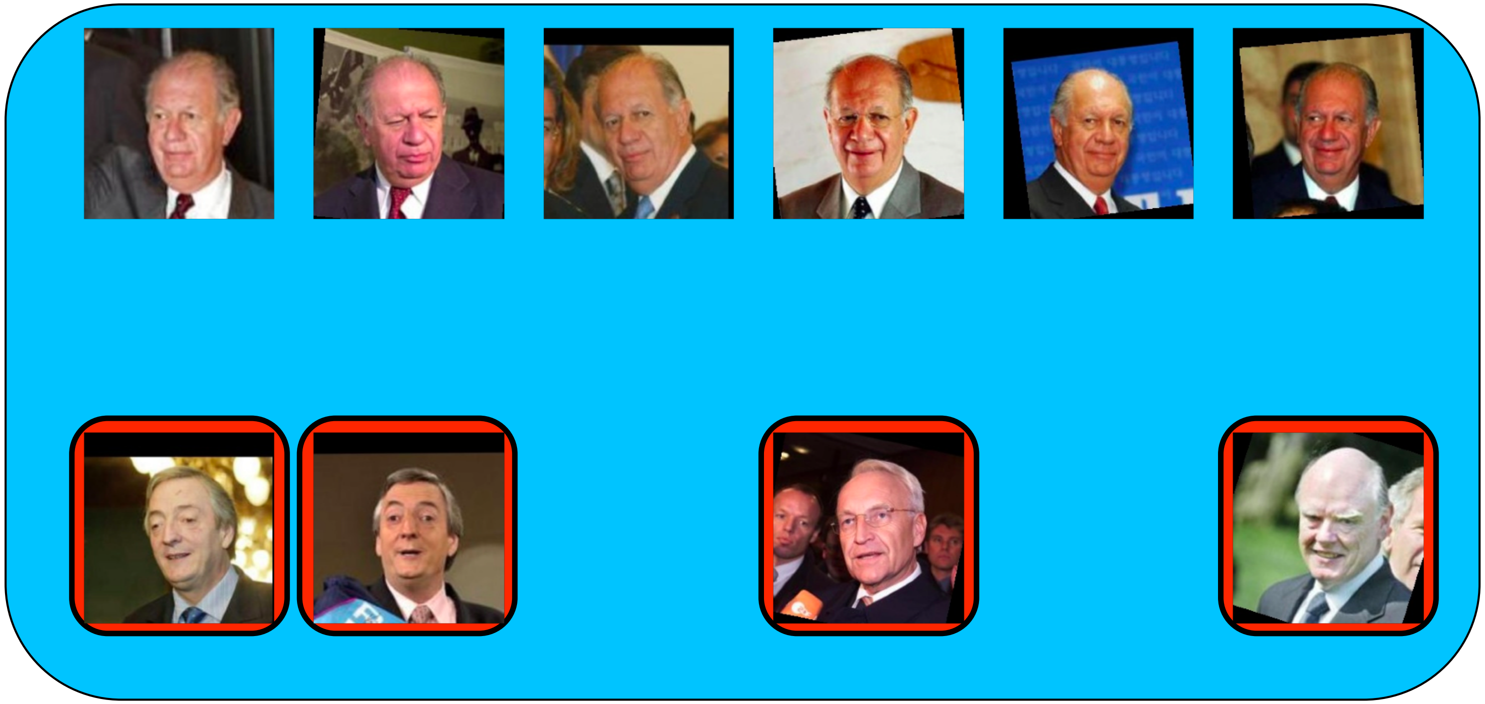
- Suppose user has two buttons
  - Button 1: Assign name to cluster of faces
  - Button 2: Assign name to a single face

- 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)

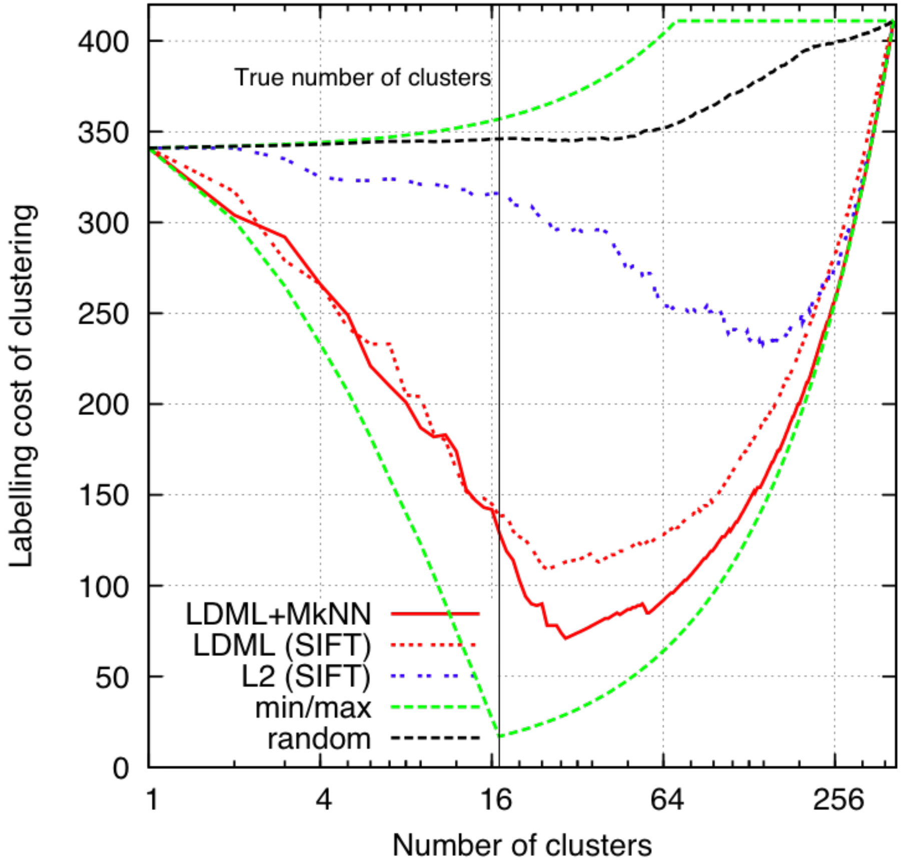
Face Clustering experiment

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



Face Clustering experiment

- Hierarchical clustering based on L2 or learned metrics
- Also compared to random clustering, min/max labelling cost
- Clustering 411 faces of 17 people
- Learned metrics yield significantly better clustering results



Example Clusters

Overview

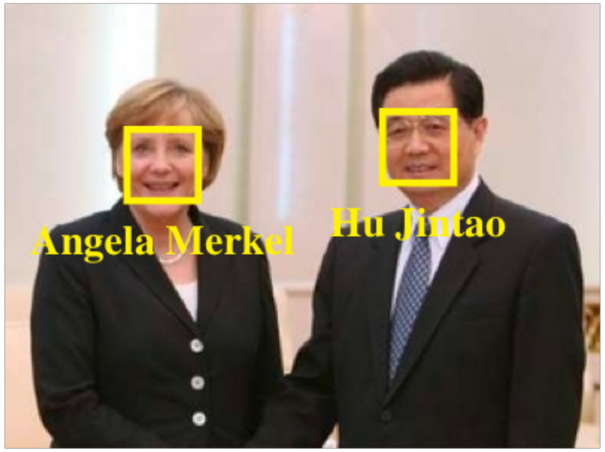
1. Metric learning methods

2. Metric learning for image annotation

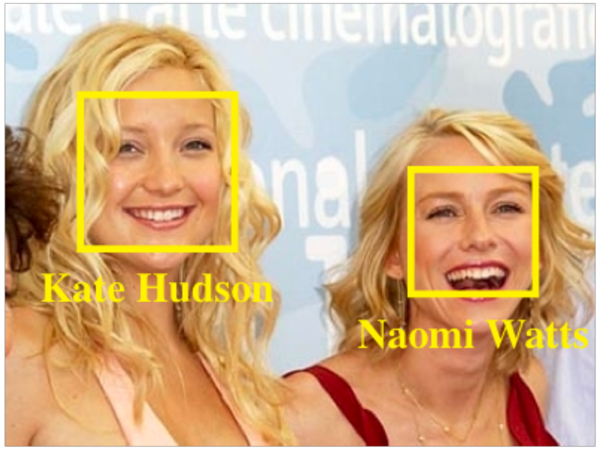
3. Metric learning for face identification
   - Application to face clustering
   - **Application to caption-based recognition**

Application 2: Caption-based recognition

- Recognition 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** (. . .)



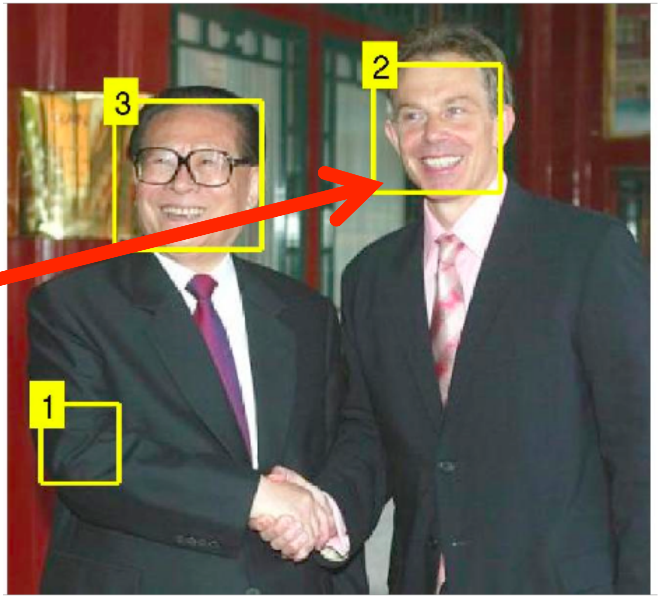
Kate Hudson and **Naomi Watts**, Le Divorce, Venice Film Festival - 8/31/2003.

- 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

Caption-based face recognition

- Iteratively optimize name-face matching per image, keeping rest fixed
- Assumptions 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 appearances with a MoG
- The discarded faces all modelled with a single Gaussian

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

- 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

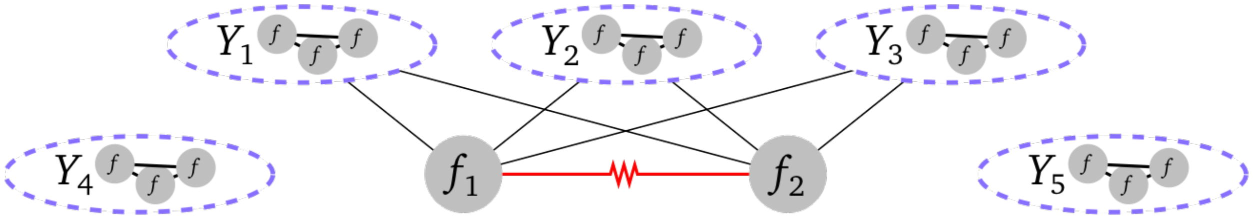
Direct similarity-based approach

- Maximize the sum of similarities between faces assigned to same name
- Fixed cost to discard a face

[Guillaumin et al. 2008]

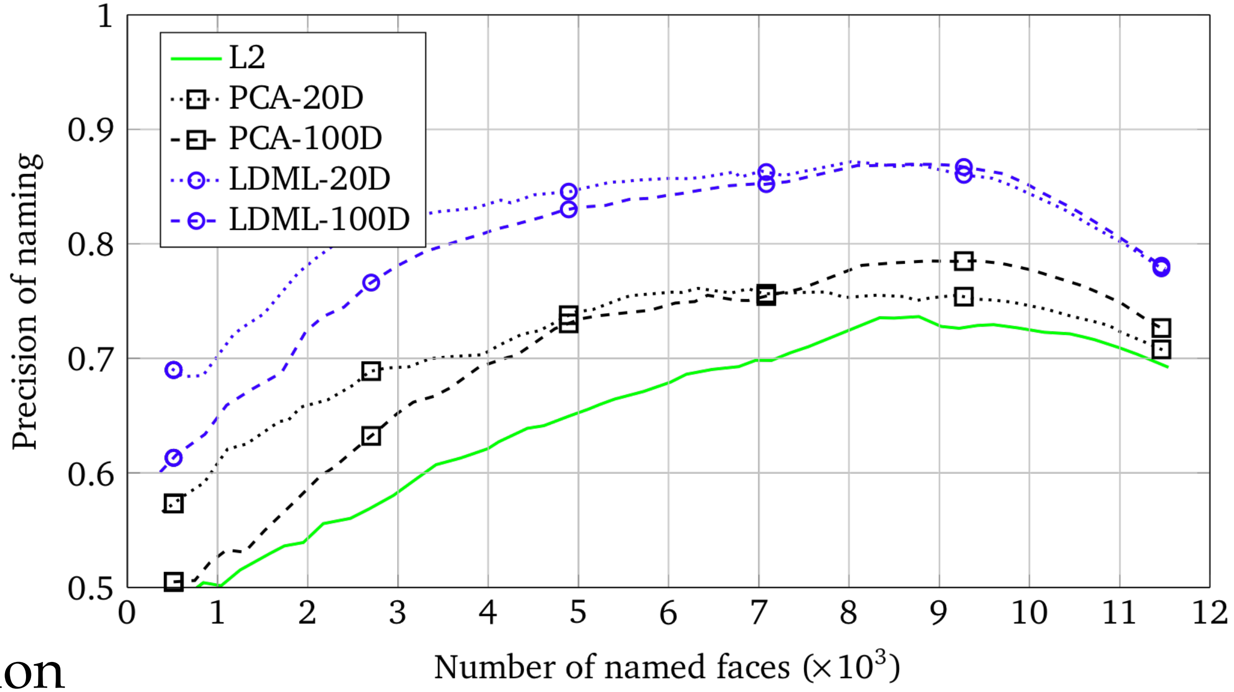
$$L(\{Y_n\}) = \sum_{n} \sum_{i \in Y_n} \sum_{j \in Y_n} w_{ij} + cN_{\varnothing}$$

- 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

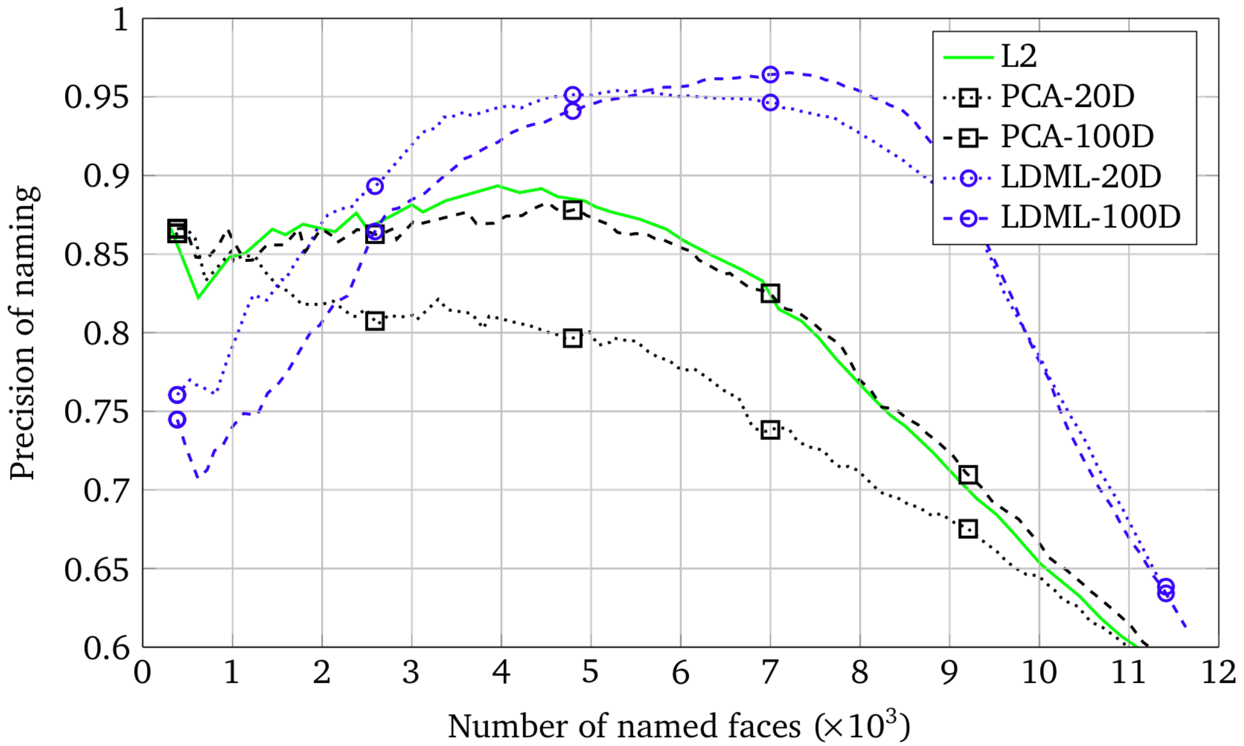
- Comparing mixtures learned in
  - Original space (L2)
  - PCA projection
  - LDML projection



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

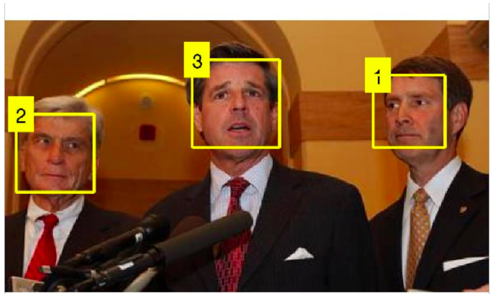
Caption-based recognition similarity-based

- Weights defined using distance from L2, PCA, LDML

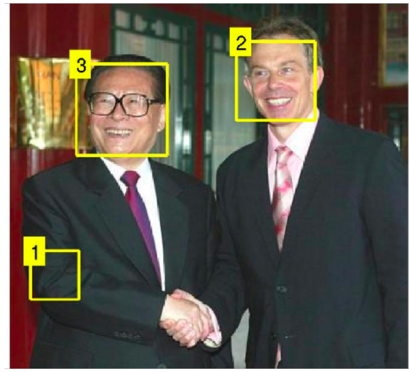


- PCA does not help: it preserves distances
- LDML: distances emphasise variations relevant for identity

Example name-face associations

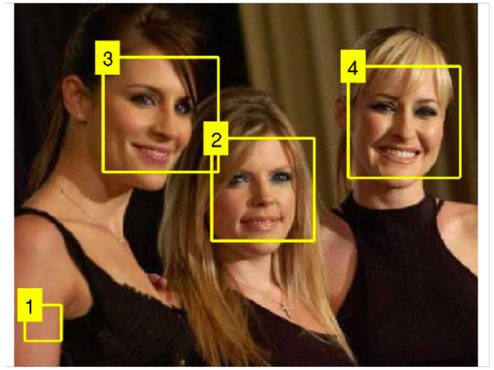


LDML	1. Saddam Hussein
	2. **John Warner**
	3. **Paul Bremer**
PCA	1. **Bill Frist**
	2. Paul Bremer
	3. Saddam Hussein



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

Take-home message

- Measures of distance or similarity appear in many places in vision
- Features of descriptors are often quite generic
- It pays-off to learn the right similarity measure for your task

References

- E. Nowak and F. Jurie. Learning visual similarity measures for comparing never seen objects. CVPR 2007.
- C. Bishop. Pattern Recognition and Machine Learning. Springer, 2006.
- L. Yang and R. Jin. Distance metric learning: A comprehensive survey. Technical report, Department of Computer Science & Engineering, Michigan State University, 2006.
- M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. ICCV 2009.
- M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? Metric learning approaches for face identification. ICCV 2009.
- M. Everingham, J. Sivic, and A. Zisserman. 'Hello! My name is... Buffy' - automatic naming of characters in TV video. BMVC 2006.
- T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. Learned- Miller, and D. Forsyth. Names and faces in the news. CVPR 2004.
- M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Automatic face naming with caption-based supervision. CVPR 2008.