Face representation and metric learning

New technologies and interfaces
for forensic face recognition
workshop EAFS 2015, Prague

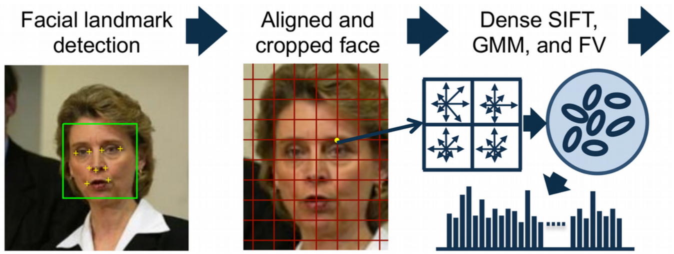
Jakob Verbeek

LEAR Team, INRIA, Grenoble, France

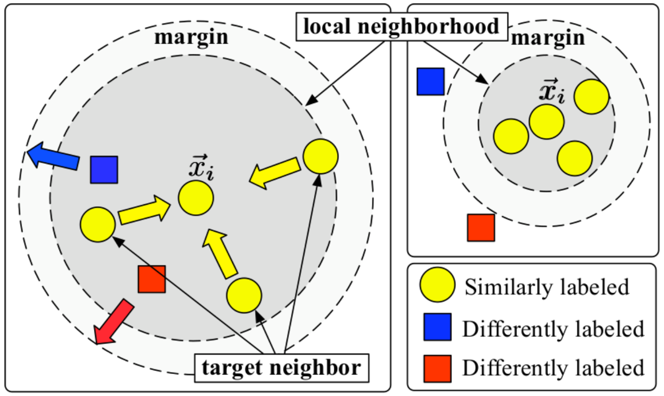
French National Computer Science Institute

Overview of the presentation

- Face representation
  - Using facial landmarks
  - Aggregated low-level statistics
  - Convolutional networks
  - Comparison



- Metric learning
  - Mahalanobis distances
  - Hierarchical metric learning
  - Local metric learning



- Age estimation

- Conclusion

Face (identity) related tasks

- Face Recognition
  - System has dataset with one or more images per person
  - Assign new face to one of these known people (or reject)

- Face Verification
  - Are two given faces of the same person or not ?
  - Should work for “new people” not seen before by system

- Face Retrieval
  - Given query face, find images of the same person in data set
  - Ranked list of results

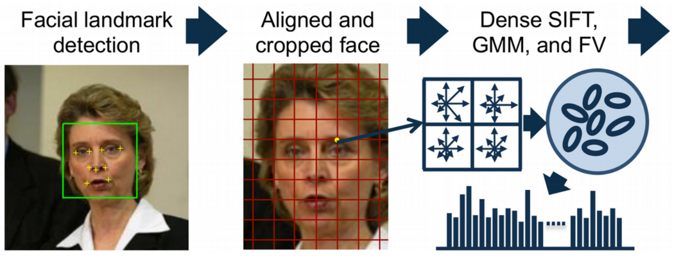
- Age estimation
- Gender, ethnicity estimation
- ….

Metric learning

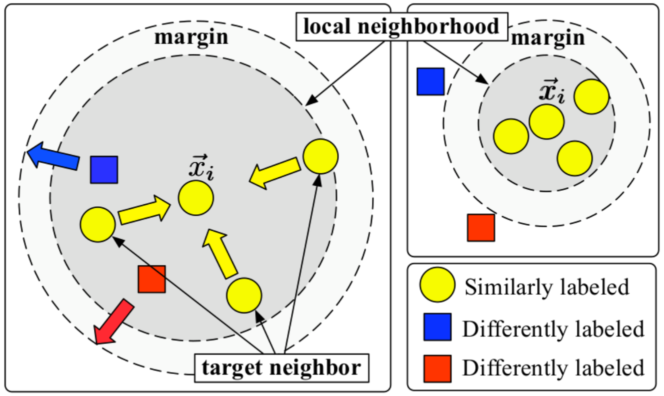
- Acquisition of measures of distance or similarity from examples
- Similarity is inherently task dependent

Overview of the presentation

- Face representation
  - Using facial landmarks
  - Aggregated low-level statistics
  - Convolutional networks
  - Comparison



- Metric learning
  - Mahalanobis distances
  - Hierarchical metric learning
  - Local metric learning



- Age estimation

- Conclusion

Challenges in face representation

- In classic “controlled” data sets nuisance factors are controlled
  - Illumination, pose, expression
  - Cooperative subjects

- Example images from the “Multi-PIE” dataset

Challenges in face representation

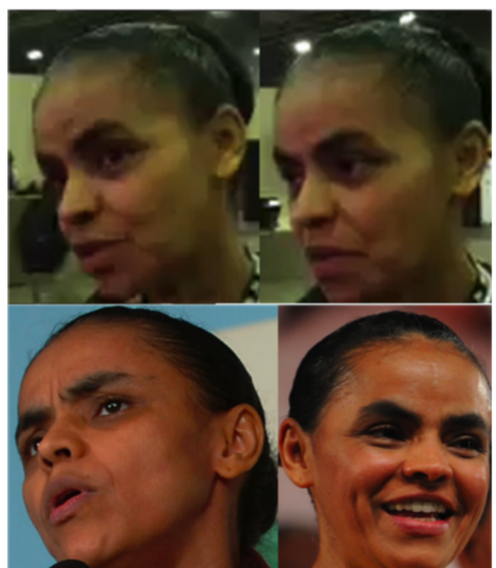
- Recent shift of attention towards “uncontrolled” datasets
  - Richer variations in nuisance factors: occlusion, illumination, expression, hairstyle, pose, *etc.*
  - Data *collected* instead of *generated* for research purposes
    - Typically collected from the web

- Examples images from the “Labeled Faces in the Wild” dataset (left) ECCV'08 and IARPA “Janus” dataset (right), CVPR 2015.

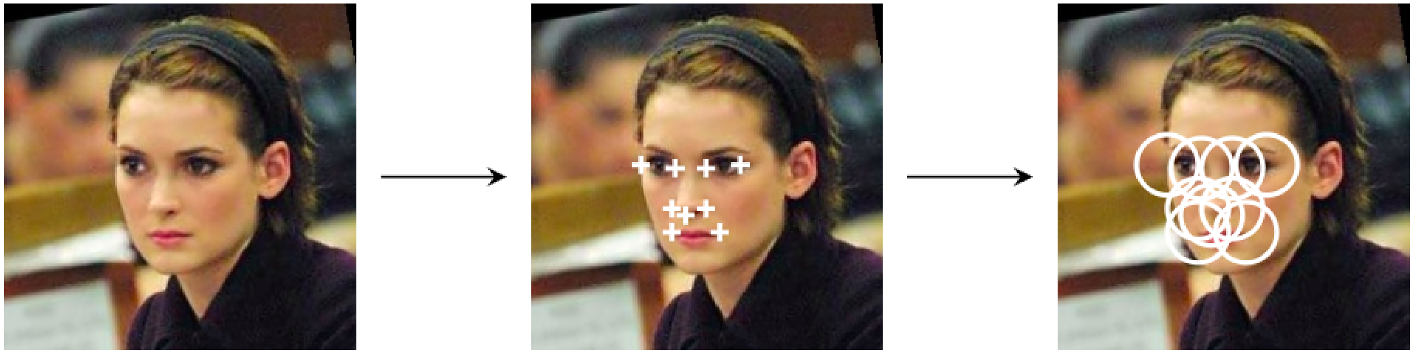
Challenges in face representation

- Desiderata of a “good” face representation
  - Efficient to compute, small memory footprint
  - Invariant to nuisance factors, effective for a range of tasks

- Sparse landmark-based approach
- Dense unsupervised local feature approach
- Dense supervised feature learning


Landmark-based face representation

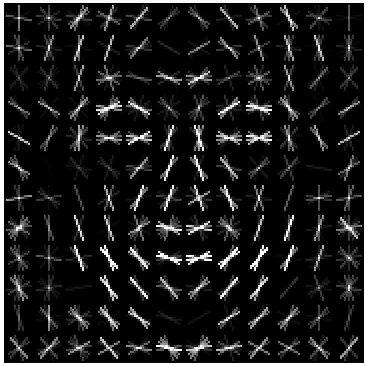


- Represent face with local descriptors of landmarks
  - Everingham et al., BMVC 2006
  - Landmarks: point on eyes, nose, mouth, ...

- Detect landmarks
- Warp face image to correct for pose (translation, rotation, scaling)
- Represent each landmark using local descriptor
  - Ignore position of landmarks in signature

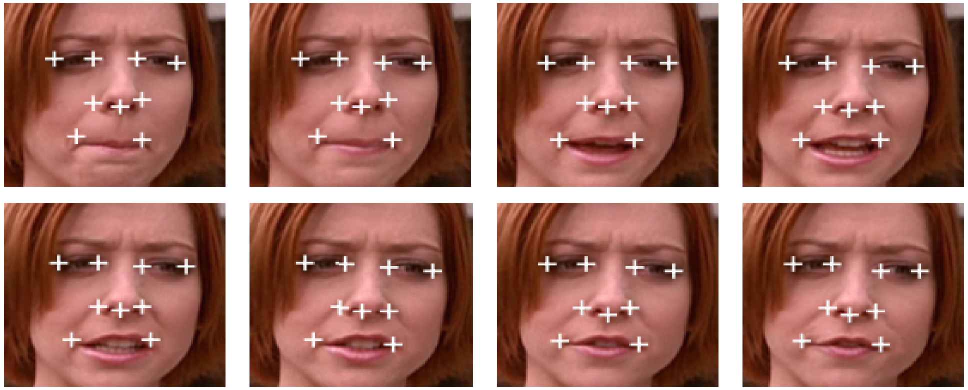
Landmark detection with constellation models

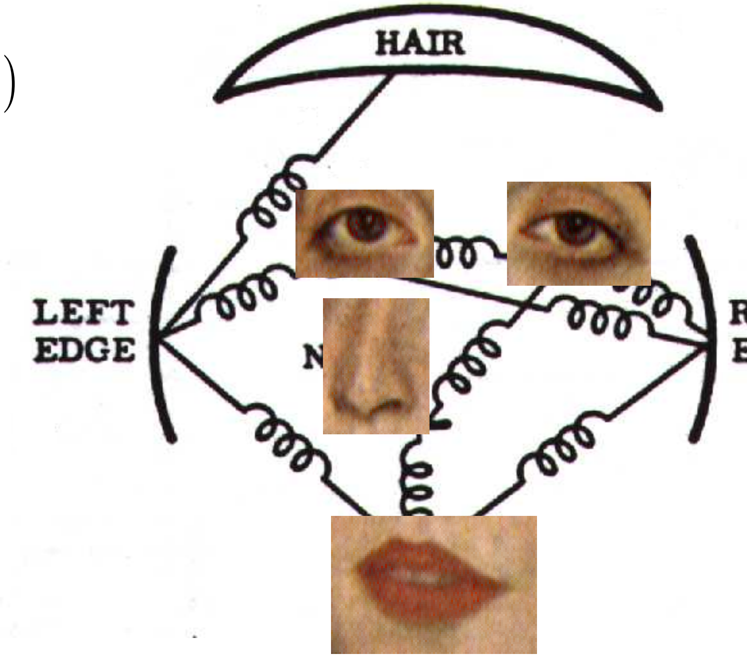
- Separate detectors for 9 facial landmarks
  - Linear HOG classifiers, Dalal & Triggs, CVPR 2005
  - Response/score map for each landmark



- Combine with displacement model between landmarks
  - Felzenszwalb & Huttenlocher, IJCV'05

$$E(x_{1,}\ldots x_9)=\sum_{i=1}^{9} S_i(x_i)+\sum_{i=2}^{9} D_i(x_i, x_{\pi(i)})$$

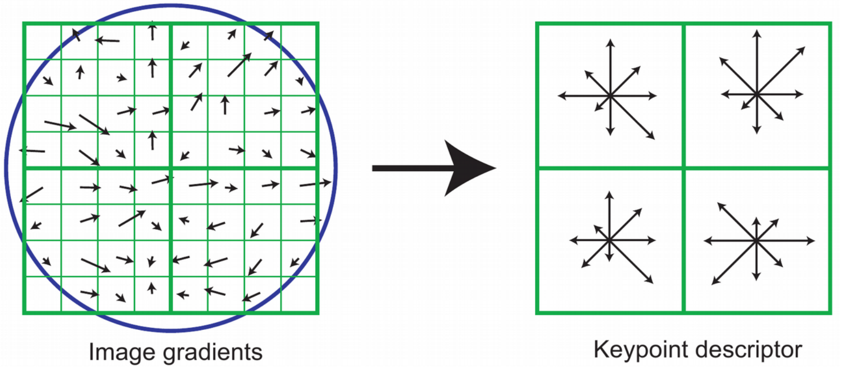




Landmark-centered feature extraction

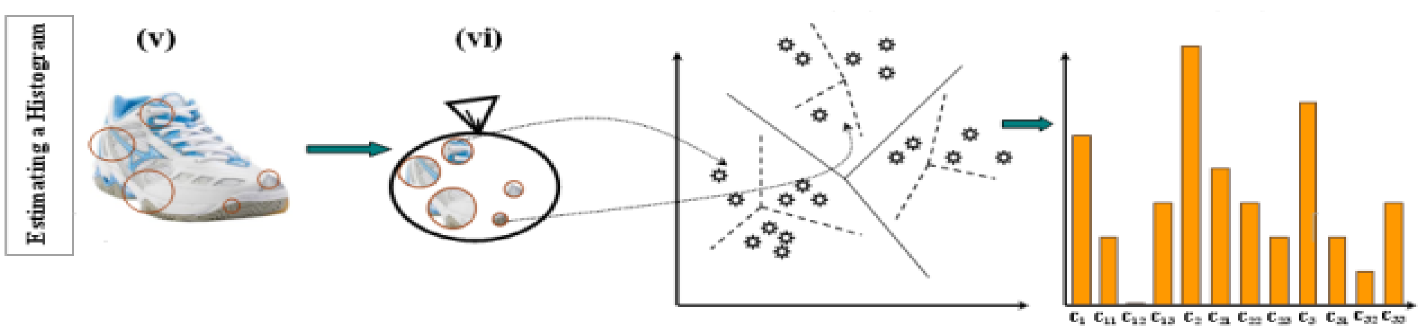
- Crop image regions around landmarks (9 landmarks, 3 scales)

- Compute 128D SIFT gradient orientation histograms (Lowe, IJCV'04)
  - Concatenate in 128 x 3 x 9 = 3,456D vector
  - Guillaumin et al., ICCV 2009



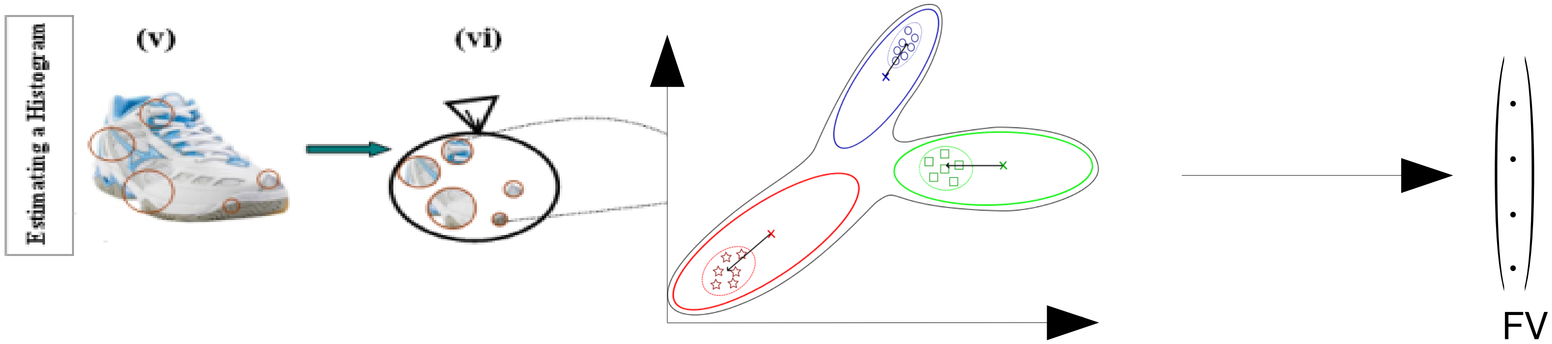
Bag-of-visual-word image representation

- Interest point detection and local descriptors (eg SIFT) have proven extremely effective for general object detection and image retrieval
  - Viewpoint invariance and robustness to partial occlusion
- Bag-of-visual-word representation
  - Sivic & Zisserman, ICCV 2003, Csurka et al. ECCV 2004
  - Cluster descriptor space to obtain discrete representation
  - Aggregate descriptors into visual word count histogram



Fisher vector image representation

- Fisher Vector (FV) representation improves over bag-of-words (BOW)
  - Perronnin et al., ECCV 2010
  - BOW: count descriptors per cluster
  - FV: compute first and second order moments per cluster



  - Gaussian mixture model (GMM) clustering instead of k-means

  - BOW: K dimensional for K clusters
  - Fisher vector: 2KD dimensional for K clusters (typically D=64)

Image representation by aggregated local descriptors 1

- Densely sampled patches of 3x3 pixels
  - Sharma et al., ECCV'12
  - Subtract value of center pixel for illumination invariance
  - Face represented by “point-cloud” in 8d space
  - Characterize face using Fisher vector of this point cloud
  - Concatenate descriptors computed over different face regions

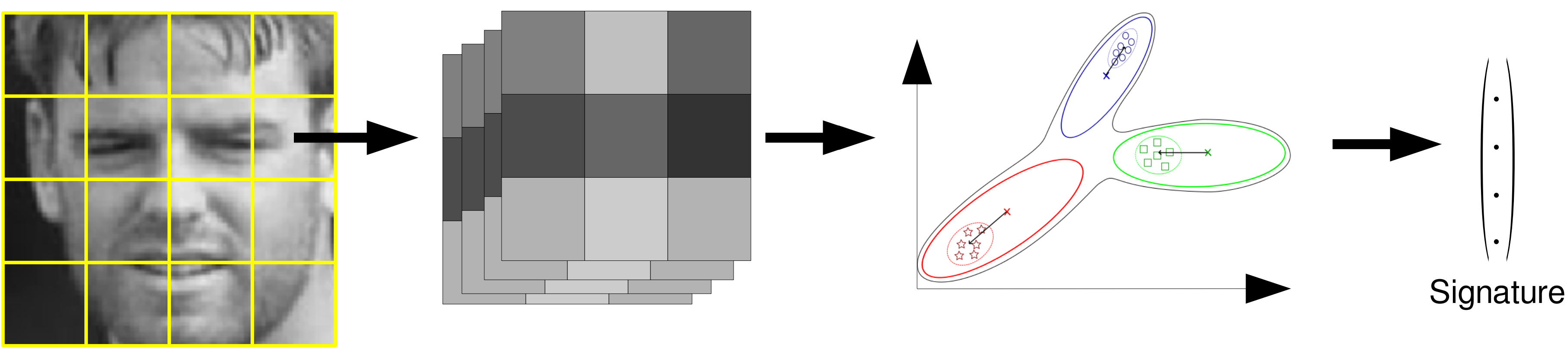
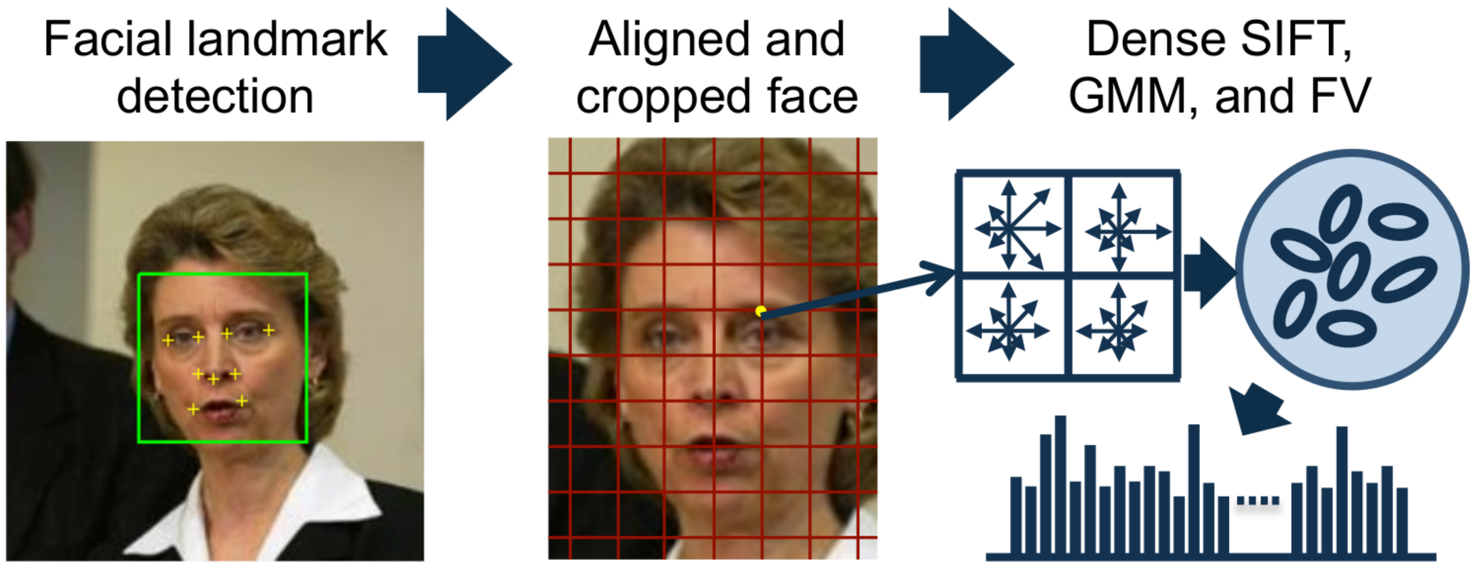


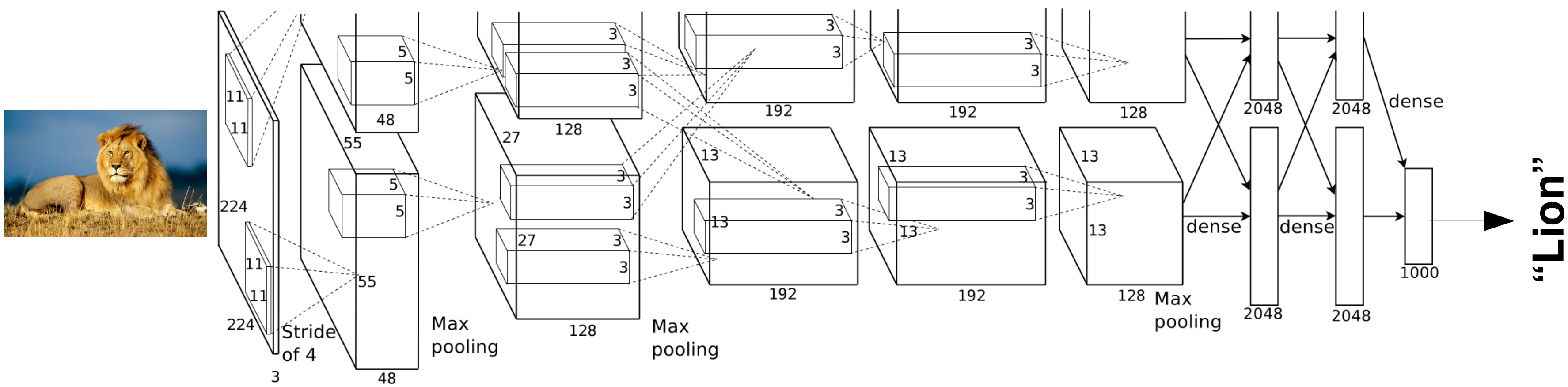
Image representation by aggregated local descriptors 2

- Densely sampled patches encoded with SIFT descriptors
  - Simonyan et al., BMVC'13
  - Concatenate 2d location of patches to SIFT descriptor
  - Fisher vector computed over point cloud of expanded descriptors



Convolutional neural networks (CNNs)

- Layered architecture of simple non-linear computations
- First computations start directly from image pixels
- End-to-end learning: Large set of parameters directly tuned to maximize performance
- Lots of success in computer vision since 2012 ImageNet succes
  - Krizhevsky et al, NIPS 2012, reduced error rate by one third
  - Most ideas date back two decades Le Cun et al, NIPS 1989
  - Millions of parameters, needs lots of data, training on GPU

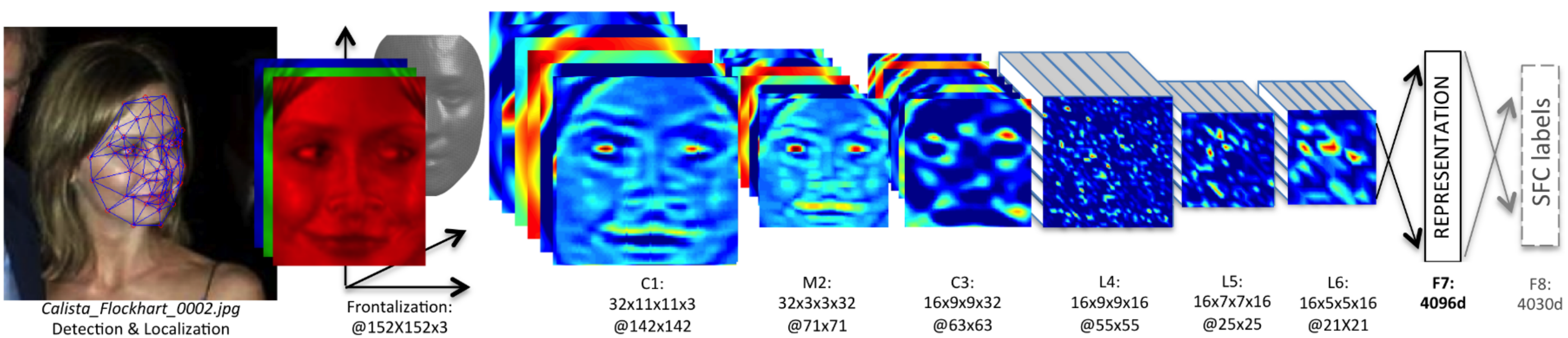


Krizhevsky et al, NIPS 2012

Face representation with convolutional networks

- Previous representations are based on
  - Land-mark detection, at least for alignment
  - “Hand-crafted” SIFT or other local features
  - Unsupervised clustering used in Fisher vectors

- Representations using convolutional neural networks
  - Often landmark-based alignment as pre-processing
  - “Hand-crafted” architecture of the network
  - Supervised learning of parameters, e.g. for face recognition

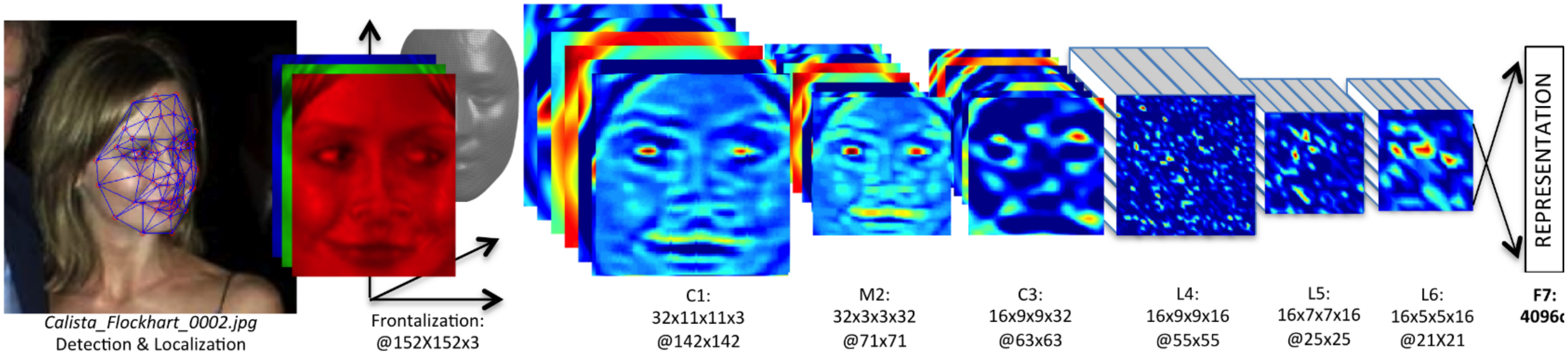


Taigman et al., CVPR 2014

Using CNN features for other tasks

- Suppose we have lots supervised data for one task, very little or no training data for another task
  - Many face images of many identities for recognition
  - Face verification for people not seen during training

- Use the “internal” representation of CNN as an image “signature”
  - Girshick et al., CVPR 2014. Taigman et al, CVPR 2014.



Taigman et al., CVPR 2014

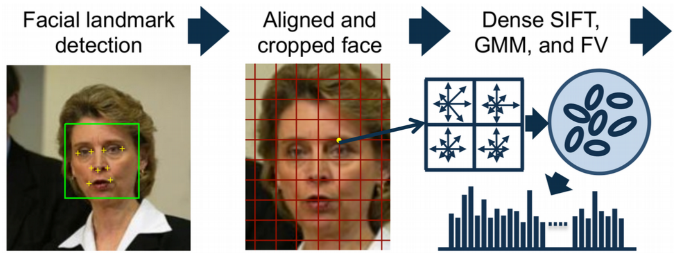
Overview of the presentation

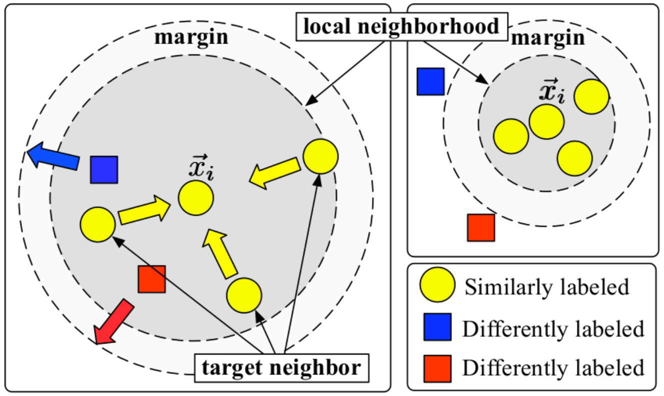
- Face representation
  - Using facial landmarks
  - Aggregated low-level statistics
  - Convolutional networks
  - Comparison

- Metric learning
  - Mahalanobis distances
  - Hierarchical metric learning
  - Local metric learning

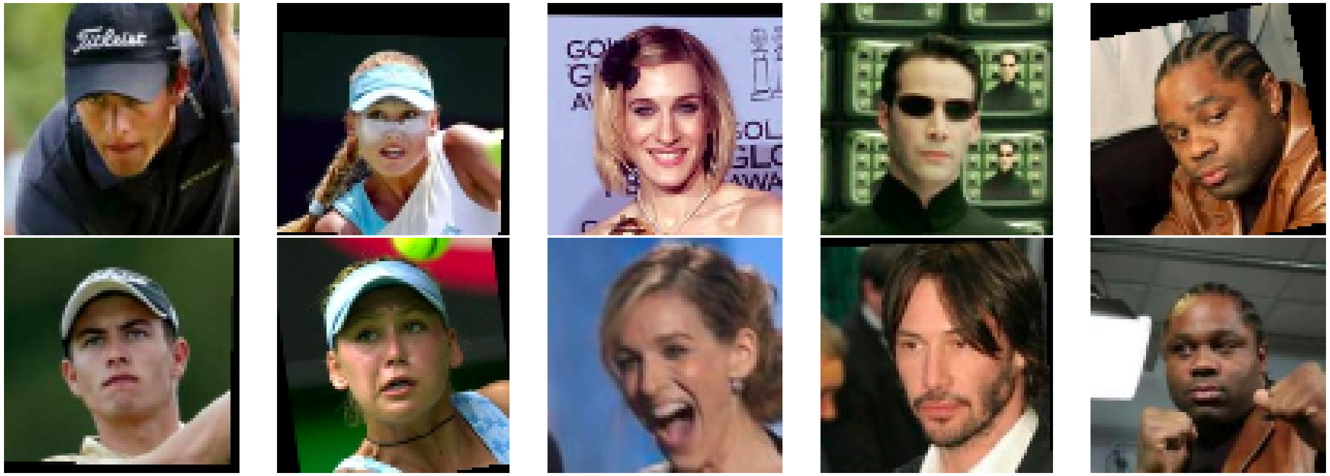
- Age estimation

- Conclusion





Experiments with “Labeled Faces in the Wild” dataset



- Contains 12,233 faces of 5,749 different people
  - http://vis-www.cs.umass.edu/lfw
- Task: given two faces, it is the same person or not ?
  - Learn metric from 90% of data, test on other 10%
  - People in test set are not in the training set
  - Performance: percentage of pairs correctly classified

Results using representations based on local features

- Performance without metric learning
  - Landmark-based SIFT approach 67.8 %
  - Fisher vector, raw 3x3 patches 73.4 %
    - Same, but our refined implementation 80.7 %
- Performance with metric learning
  - Landmark-based SIFT approach 83.2 % (+15.4)
  - Fisher vector, our optimized implementation 86.4 % (+ 5.7)
  - Fisher vector, dense SIFT 91.4 %

- Dense features improve over landmark-based ones
- Surprisingly good performance using simple 3x3 patches
- Metric learning improves performance significantly

Results using CNNs

- Local features with metric learning from 13K images
  - Fisher vector, dense SIFT , Simonyan et al, 2013 — 93.1 %
- Recent CNN-based results

Ours	500K	95.2 %
• + local metric learning	500K	96.8 %
Parkhi et al., BMVC 2015 +	2.6M	99.0 %
Taigman et al., CVPR 2014 (facebook)	4M*	97.4 %
Sun et al., CVPR 2014	200K*	97.5 %
Schroff et al., CVPR 2015 (google) +	200M*	98.9 %
• With face alignment		99.6 %

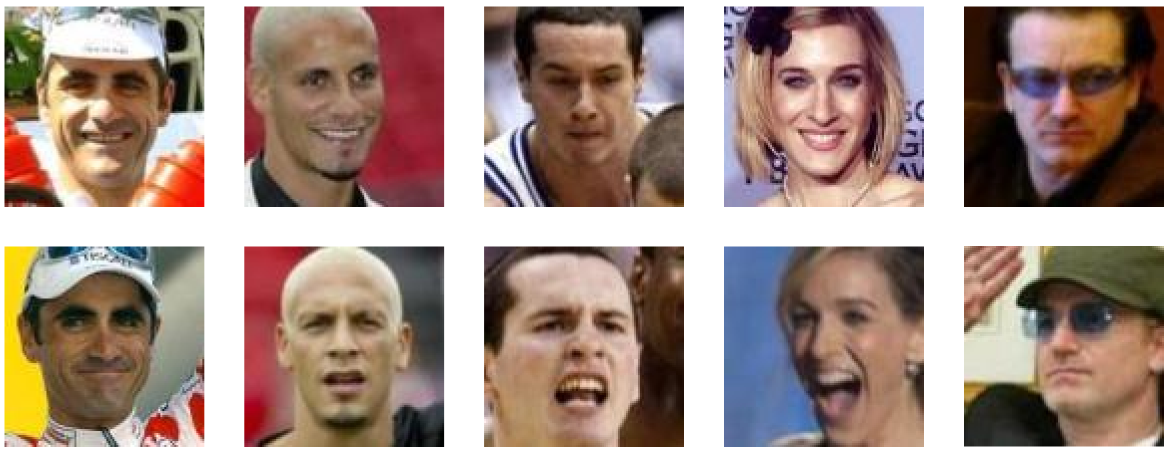
+ metric learning drives CNN training

* results based on proprietary datasets, not reproducible

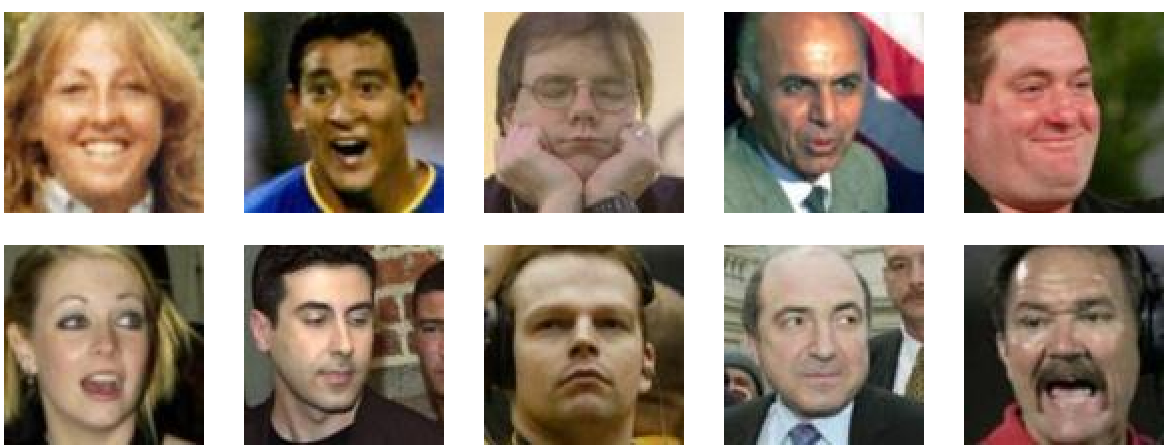
- CNN features improve results using more training data: size matters
- Best results using metric learning to drive CNN training

Hard cases: correct decision, closest to being wrong

- Same person: illumination, pose, expression, occlusion
- Different people: same gender, similar hair and age



Correctly
predicted as same



Correctly
predicted as different

Hard cases: strongest response for wrong decision

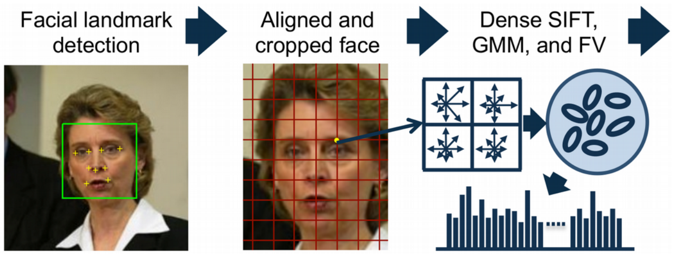
- Same person: occlusion, blur, pose
- Different people: people with same gender and ethnic background

Incorrectly predicted as different

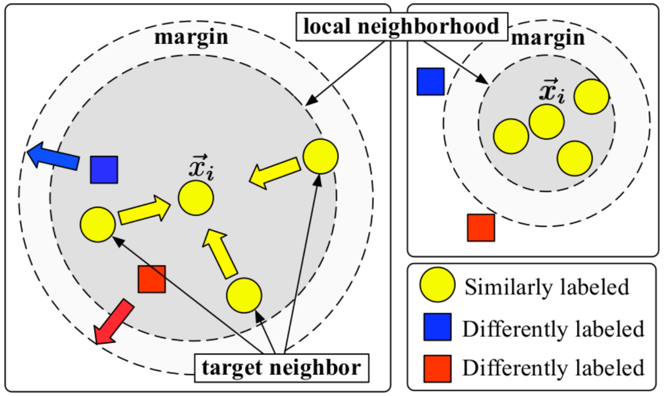
Incorrectly classified as same person

Overview of the presentation

- Face representation
  - Using facial landmarks
  - Aggregated low-level statistics
  - Convolutional networks
  - Comparison



- Metric learning
  - Mahalanobis distances
  - Hierarchical metric learning
  - Local metric learning

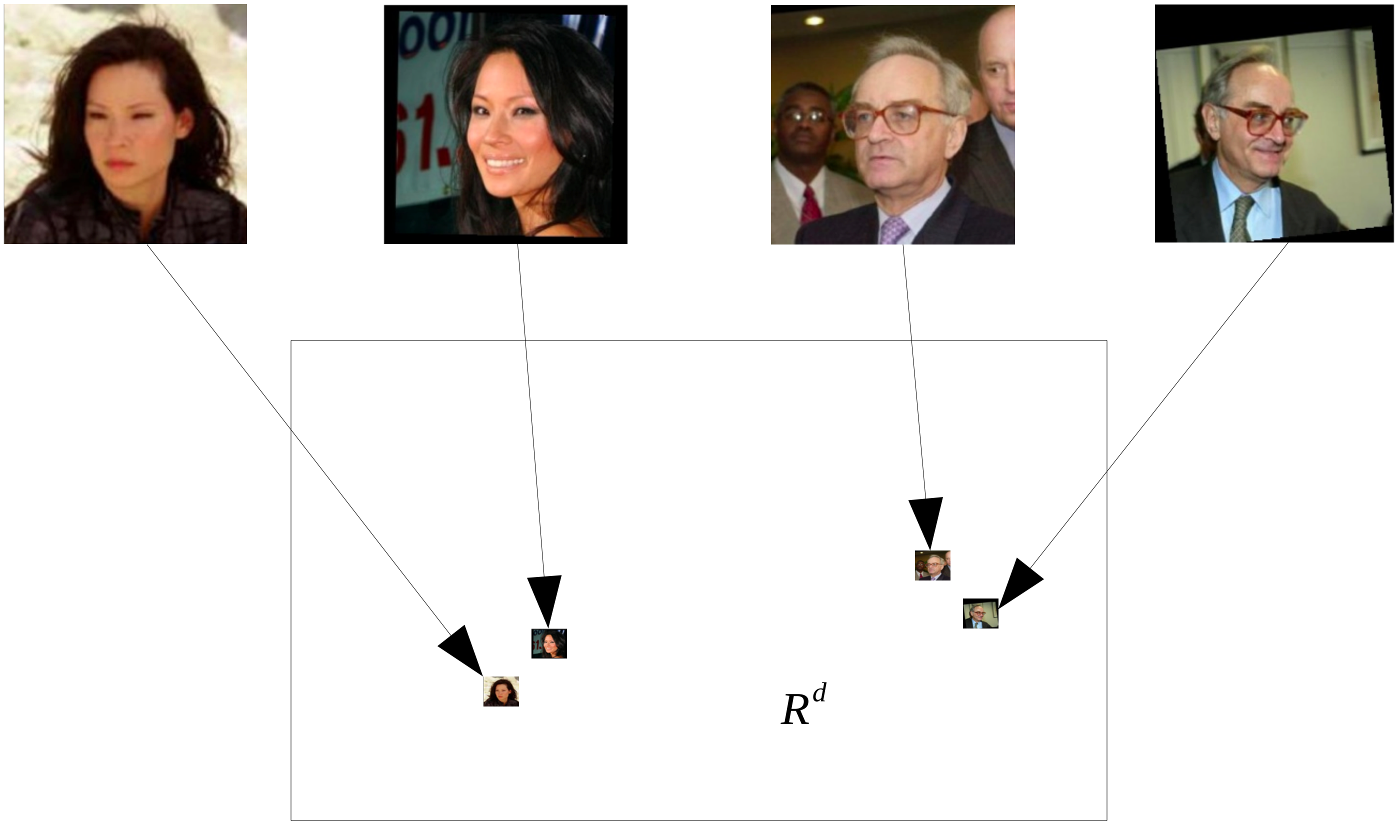


- Age estimation

- Conclusion

Metric learning

- Embed (face) given signatures in a vector space such that distance is semantically meaningful
  - Faces of same identity close, different identities far



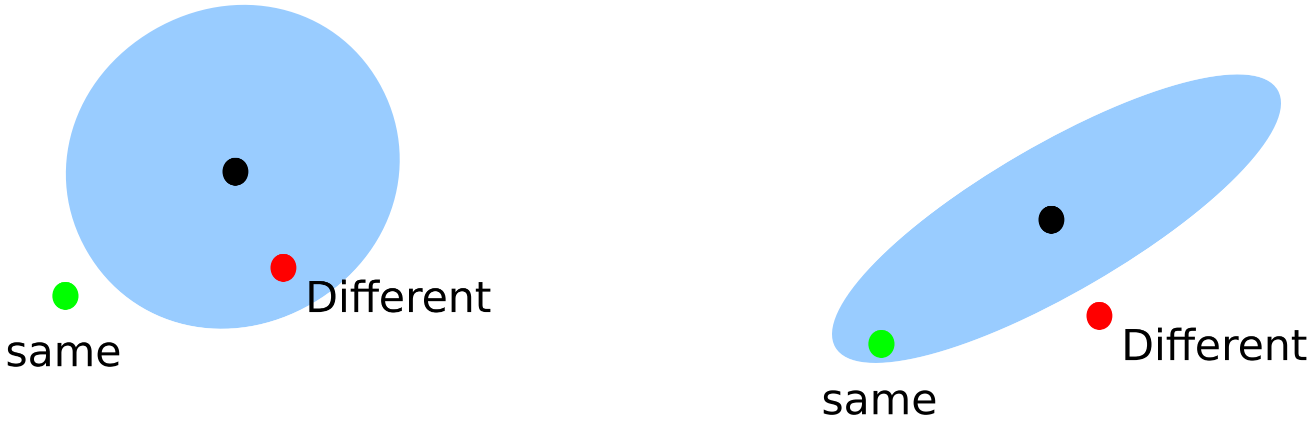
Mahalanobis metric learning

- Mahalanobis distance

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

  - Generalization of Euclidean distance: set M = I

- Equally distant points on ellipsoid instead of circle



Mahalanobis metric learning

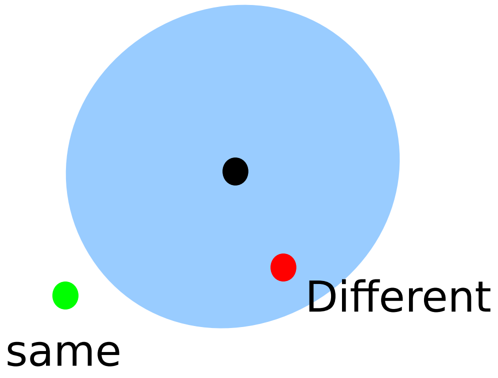
- Mahalanobis distance impractical for high dimensional data
  - Number of parameters quadratic in data dimension
  - PCA pre-processing might throw away important dimensions

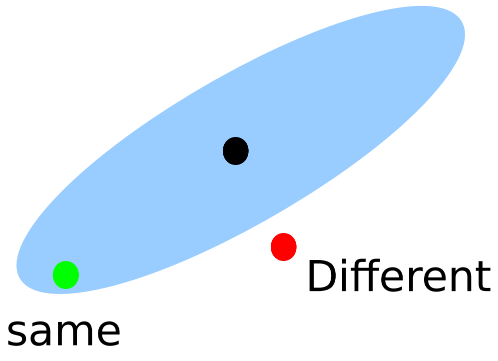
$$d_M(x, y) = (x - y)^T M (x - y)$$

- Reformulate as L2 distance after linear projection to lower dim. space

$$d_L(x, y) = (Lx - Ly)^T (Lx - Ly)$$

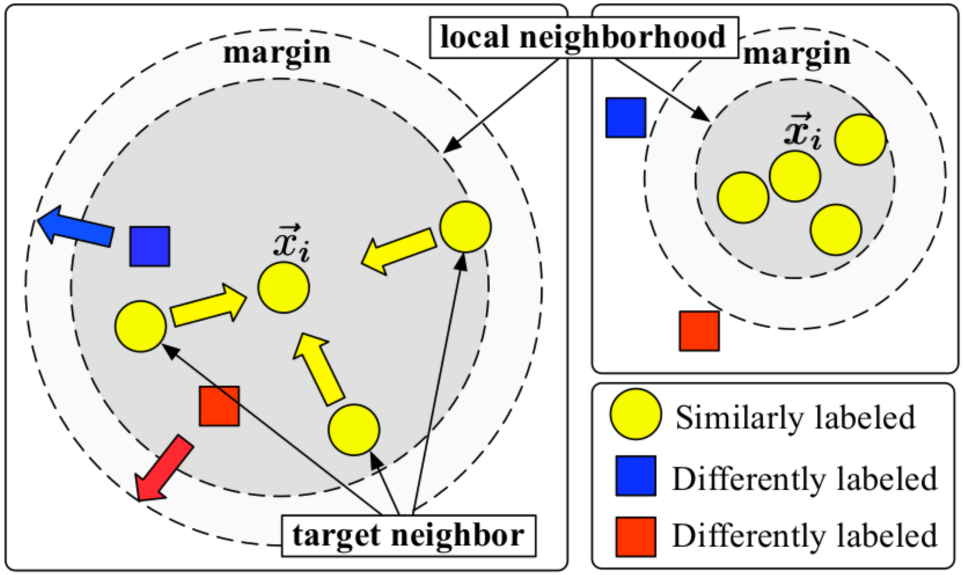
  - Number of parameters linear in data dimension
  - Can be used as data compression if L is a matrix of size $d \times D$



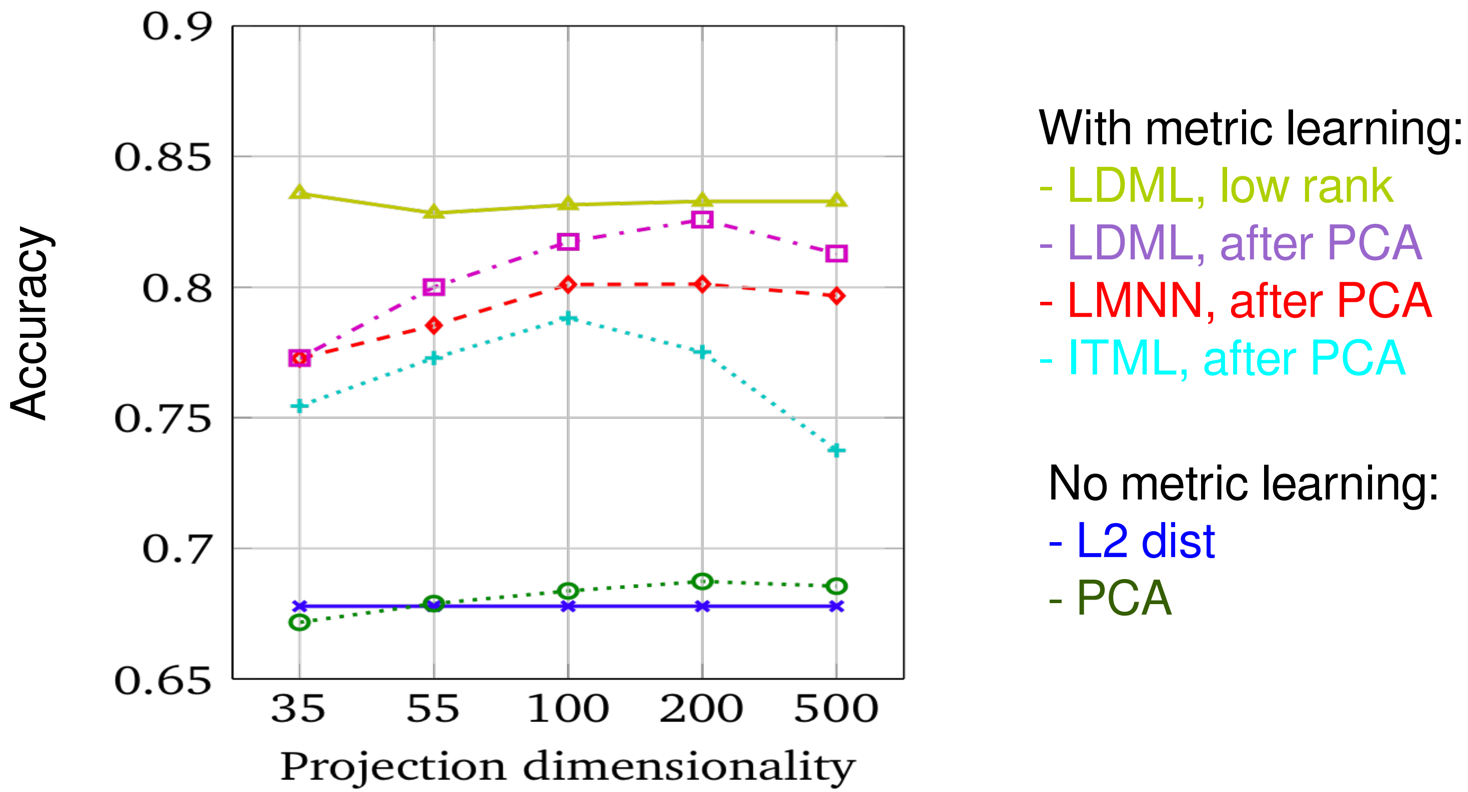


Metric learning using pairs or triplets

- Classify pairs of faces based on distance between descriptors
  - Same if $d_L(x,y) < b$ different if $d_L(x,y) \geq b$
  - Learn (L,b) using logistic discriminant classifier
  - “LDML” Guillaumin et al, ICCV 2009
- Using triplets of data points
  - Want x to be closer to y (same id) than to z (different id)
  - Triplet satisfied if $d_L(x,y) + a < d_L(x,z)$
  - “LMNN”, Weinberger et al, NIPS 2006

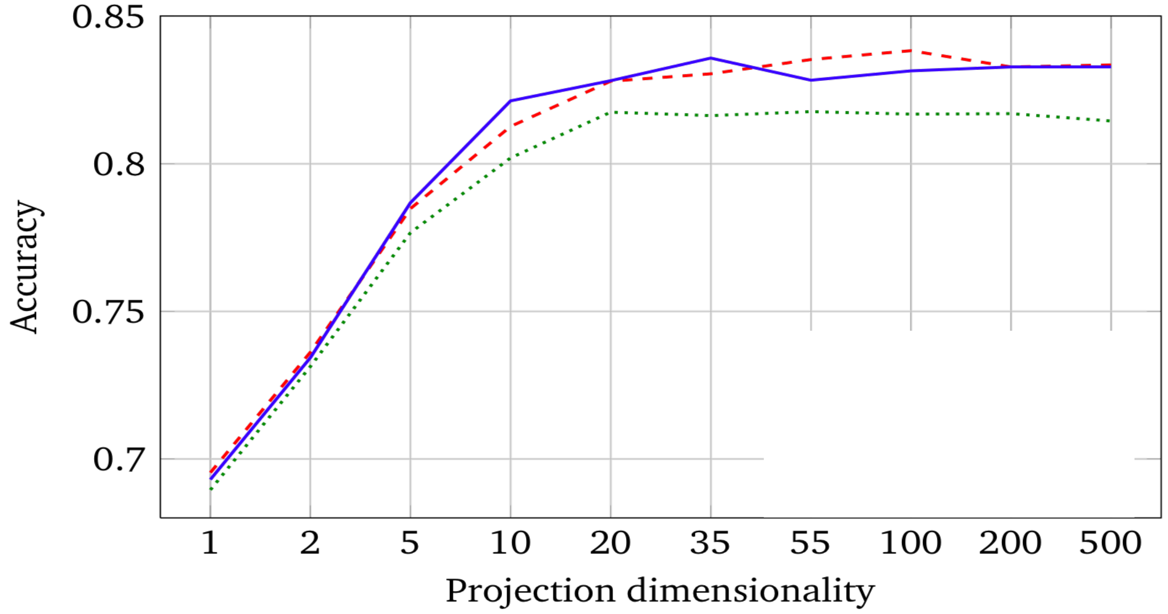


Effect of metric learning on landmark based features



- Metric learning substantially improves performance
- Low-rank metric learning better than first doing PCA
  - PCA suppresses information relevant for identity

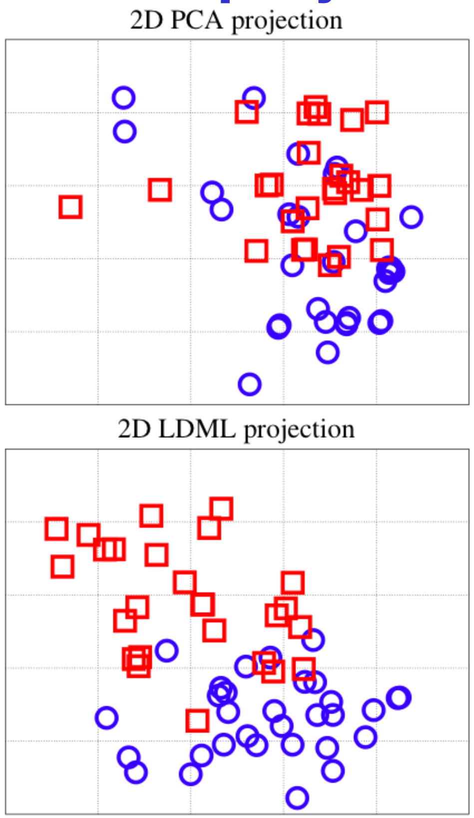
Performance as a function of projection dimension



- Surprisingly good performance with few dimensions
  - Using Euclidean distance give 67.8% correct

- Performance saturates relatively quickly
  - Original signature dimension 3,456

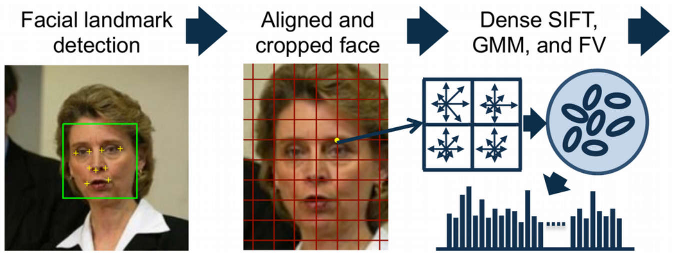
Comparing LDML and PCA projections



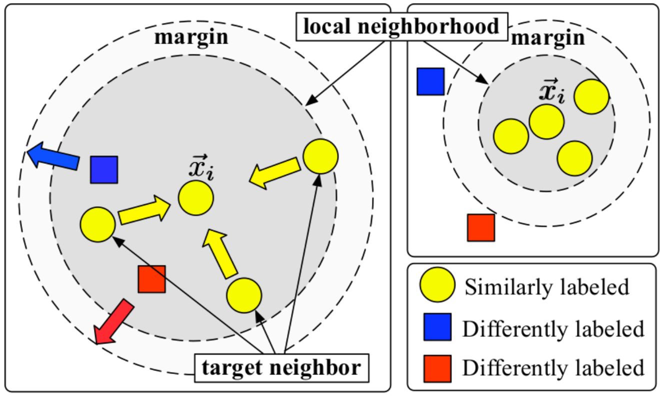
- Using PCA and LDML to find two dimensional projection of the faces of Britney Spears and Jennifer Aniston

Overview of the presentation

- Face representation
  - Using facial landmarks
  - Aggregated low-level statistics
  - Convolutional networks
  - Comparison



- Metric learning
  - Mahalanobis distances
  - Hierarchical metric learning
  - Local metric learning

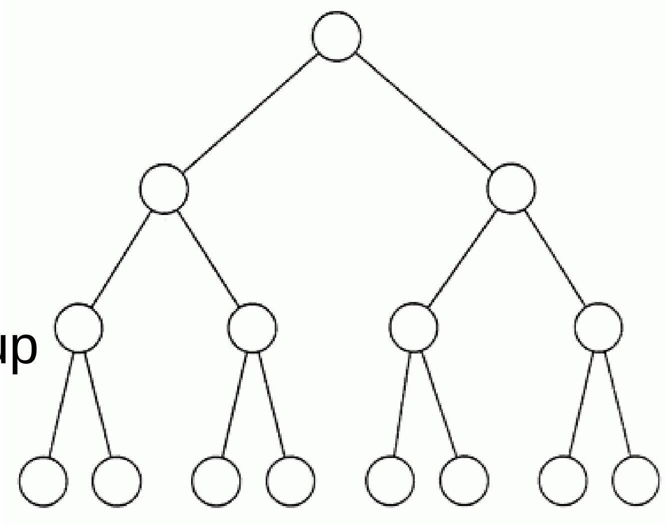


- Age estimation

- Conclusion

Hierarchical metric learning for face retrieval

- Hierarchical grouping of large face database
  - Bhattarai et al, ECCV 2014
  - Groups similar faces together
  - Assign query face to group
  - match only to faces in that group: speed-up
- Specific metrics adapted to each group
  - Important features differ per group



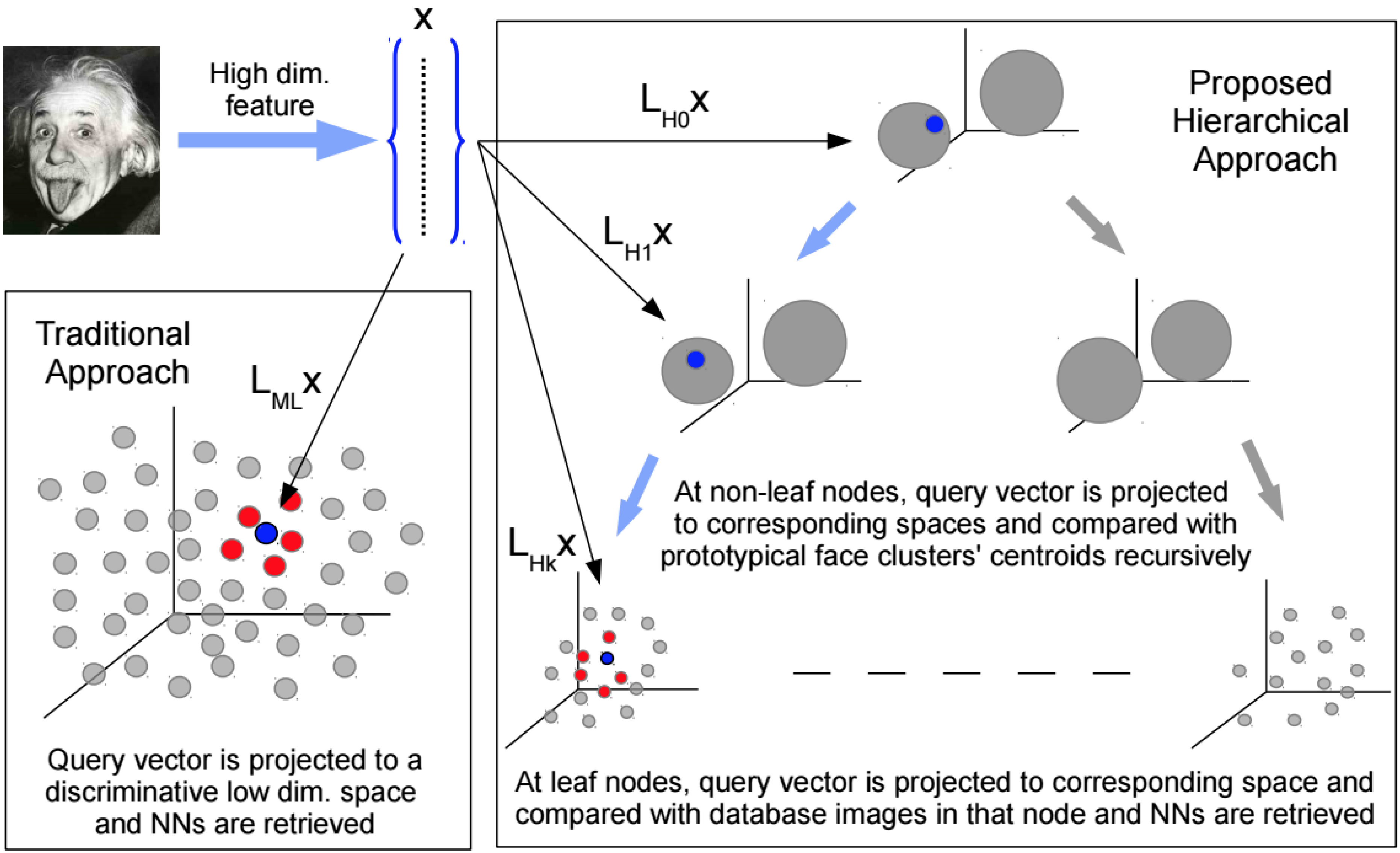
Cluster 16

Cluster 11

Cluster 15

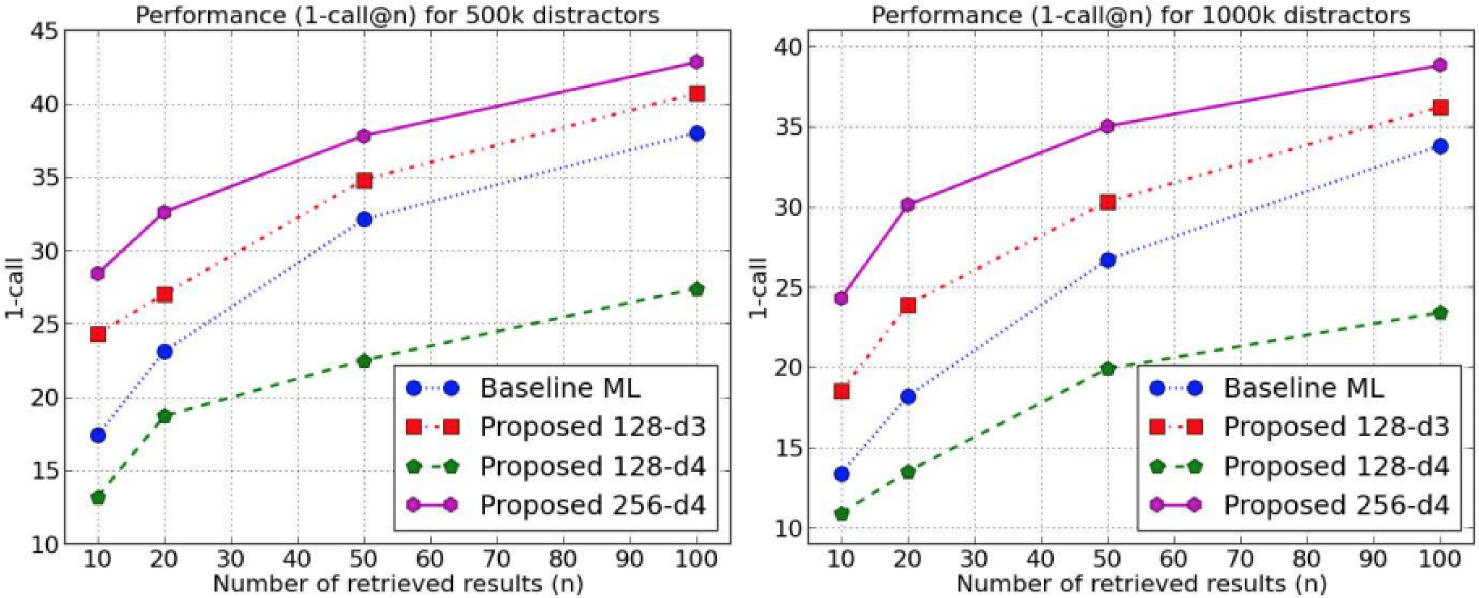
Cluster 3

Hierarchical metric learning for face retrieval: overview



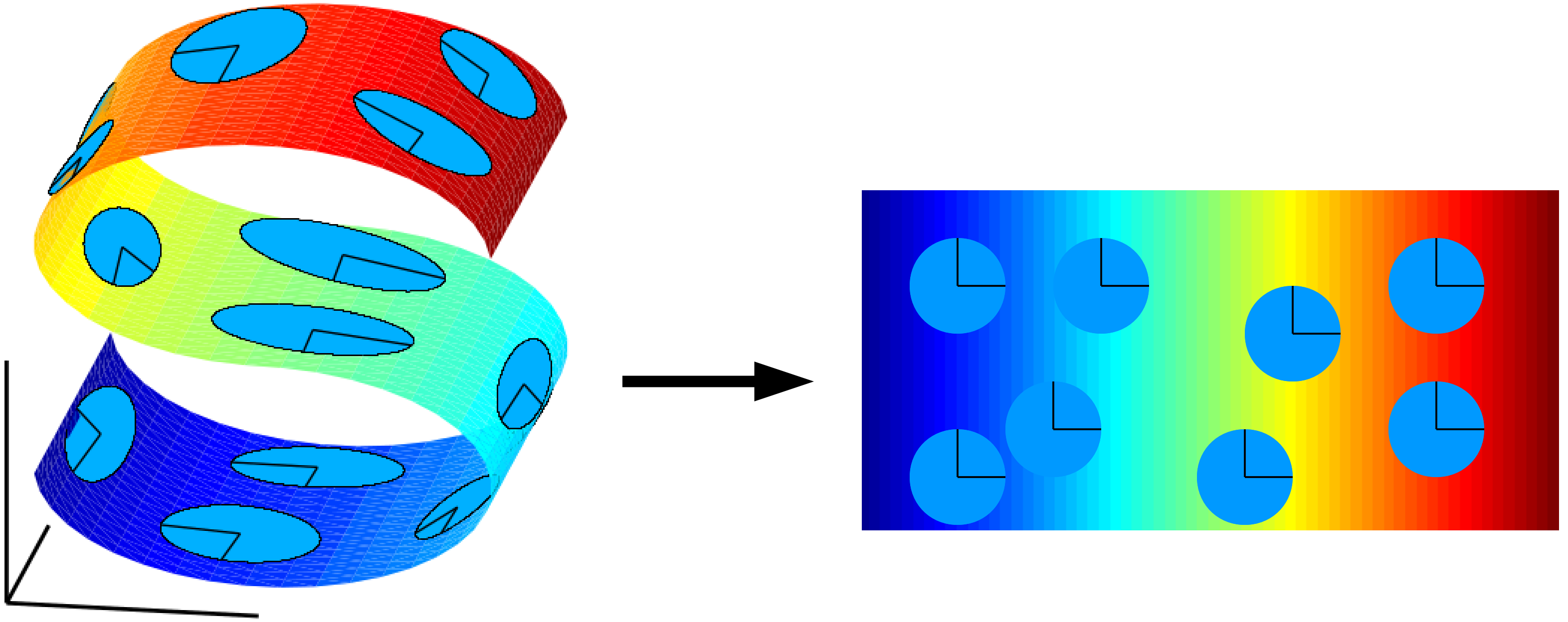
Hierarchical metric learning for face retrieval: results

- Queries from Labeled Faces in the Wild dataset
  - Additional 500,000 or 1,000,000 distractor faces added
- Performance measure: fraction of queries with correct result within the top n images
- Hierarchy can speed-up and improve results
  - d3: 8x speed-up w.r.t. baseline, d4 give 16x speedup



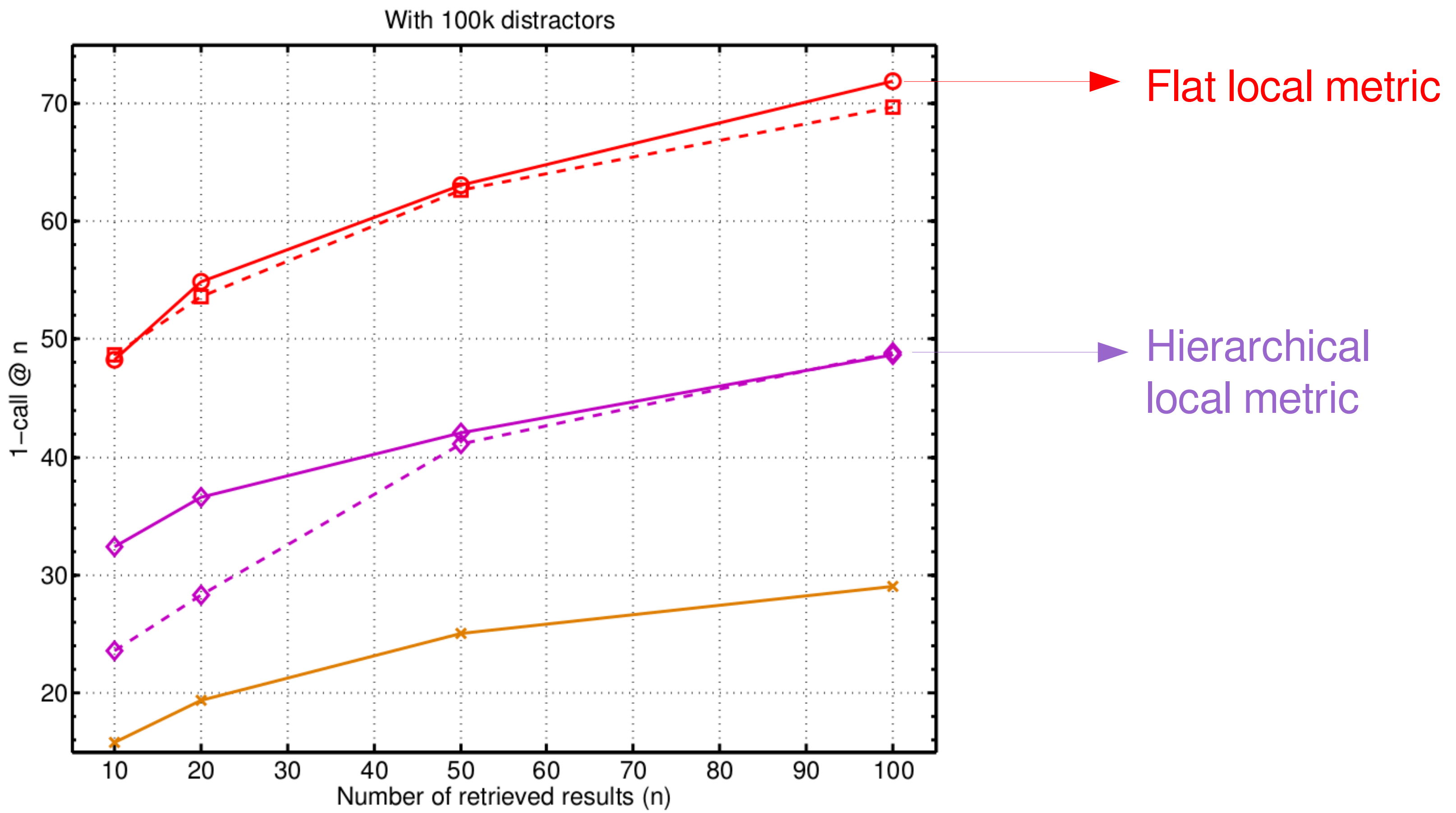
Local metric learning for face retrieval

- Grouping of large face database, learn metric per group
  - Non-hierarchical clustering avoids poor splits in top of tree
- Embed all data in a single space
  - Align local metrics via local rotations and translation
  - Can match any pair of points, not only within group



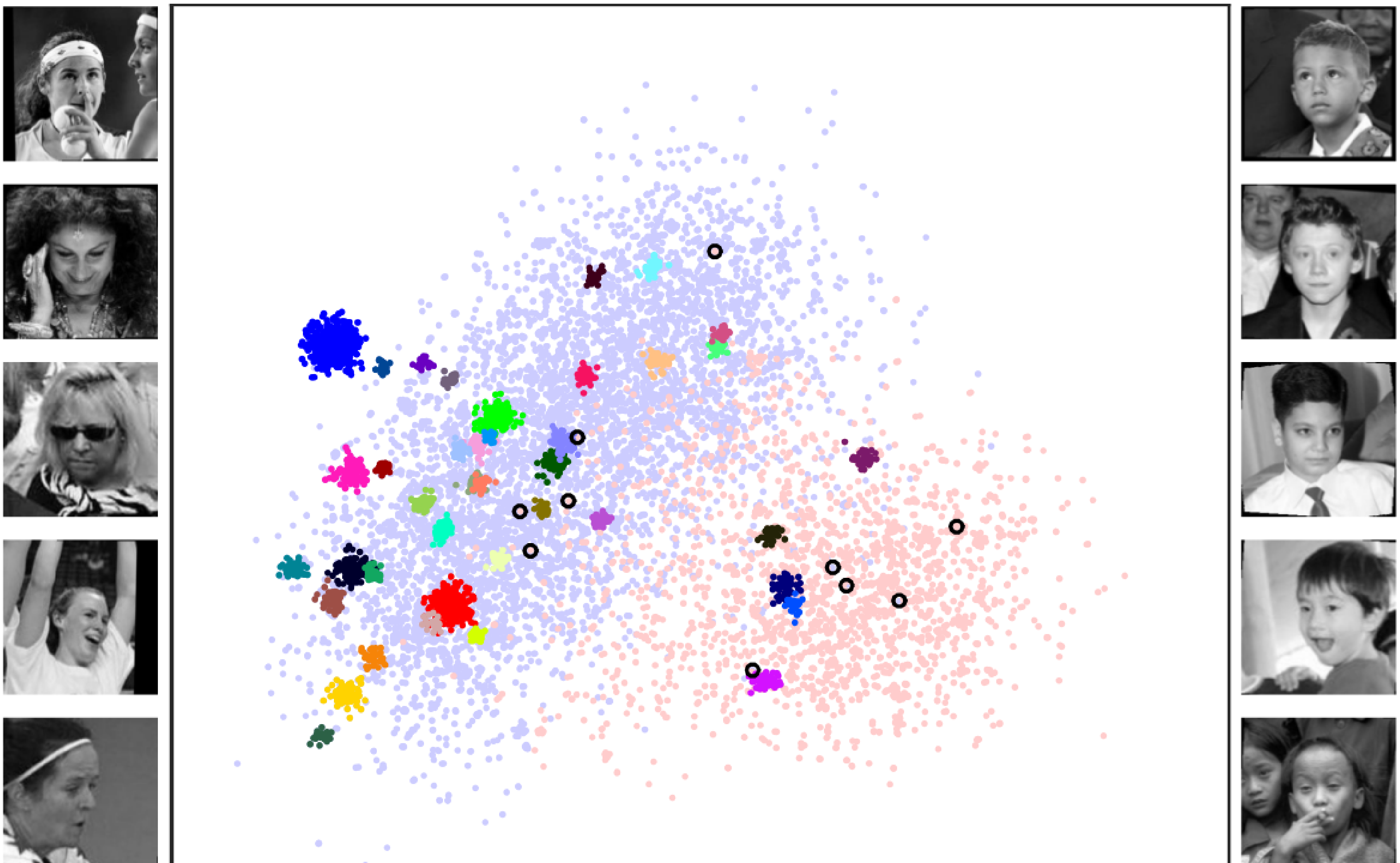
Local metric learning for face retrieval: evaluation

- Substantial improvements over hierarchical metric learning approach
  - Flat clustering more effective
  - Retrieval across full data set



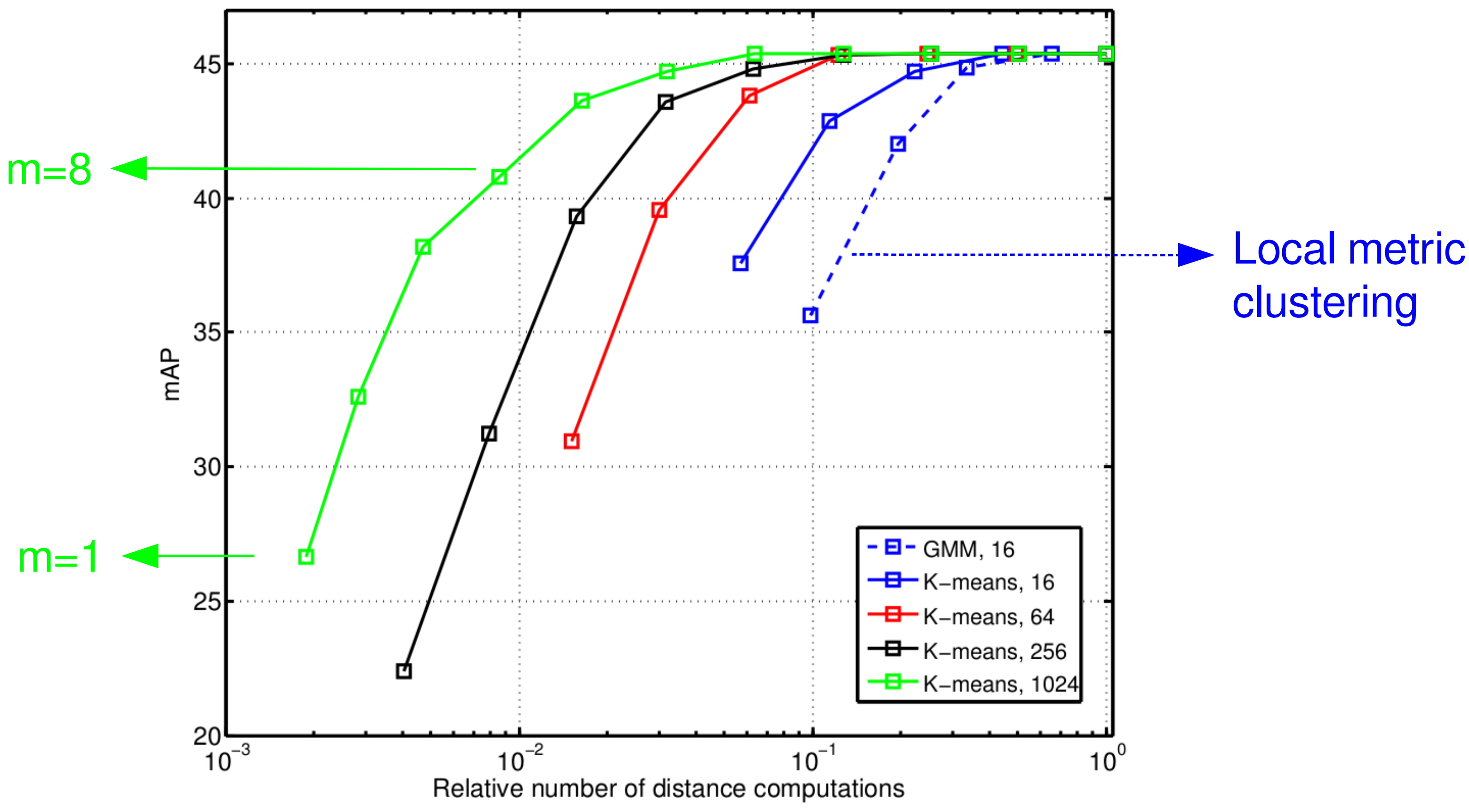
2d illustration of learned metric embedding

- Faces male/female color coded, as well as 40 people
  - Male/female separated, outliers: children and strong pose/express.



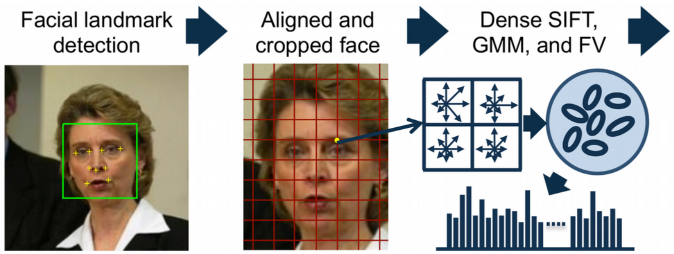
Efficient search across a dataset of a million faces

- Clustering in learned global metric embedding space
  - Match cluster of query, or the *m* nearest [Jegou et al., PAMI 2010]
- More effective than using clustering used for local metrics
  - More clusters better for any operating point

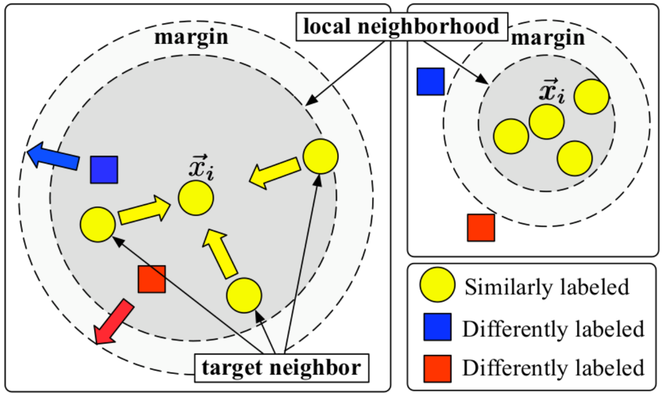


Overview of the presentation

- Face representation
  - Using facial landmarks
  - Aggregated low-level statistics
  - Convolutional networks
  - Comparison



- Metric learning
  - Mahalanobis distances
  - Hierarchical metric learning
  - Local metric learning



- Age estimation

- Conclusion

Age estimation

- Given face image predict the age of the subject: regression problem
- Aging effects differ among people from different ethnics, gender, etc.
- Training separate models per group has limitations
  - more expensive
  - very few examples in some groups

Examples from FGNET database (top row)
and the MORPH database (bottom row)

Cross-population age estimation

- Large number of training examples in “source” domains
- Few training examples in “target” domain
- Idea: Find a common linear subspace for regression
  - Source domain helps to identify subspace
  - Less regression parameters to estimate for target domain

$$\min_{L,\mathbf{w}} \mathcal{L}(\mathcal{A},\mathcal{S},\mathcal{D};L,\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \beta\sum_k \ell_{\mathbf{w}}(L\mathbf{x}_k, y_k) + \gamma \sum_{\mathcal{S}\cup\mathcal{D}} \ell_L(\mathbf{x}_i,\mathbf{x}_j,y_{ij})$$

- Regression loss: $\ell_{\mathbf{w}}(L\mathbf{x}, y) = \max(0, |\mathbf{w}^\top L\mathbf{x} - y| - \epsilon)$
- Metric learning loss based on age: $\ell_L(\mathbf{x}_i,\mathbf{x}_j,y_{ij})$
  - Using cross-domain pairs

Results on Morph II dataset

- Four domains: White Female, White Male, Black Female, Black Male
- Target size: number of training images in target domain

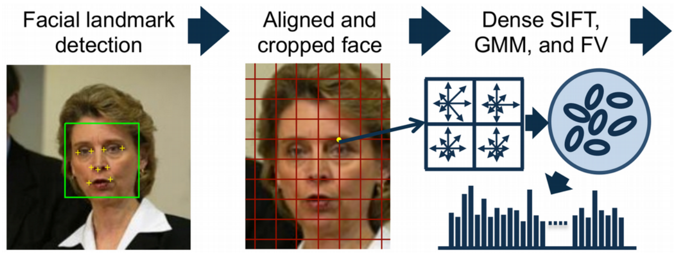
- Comparison
  - LBP: no subspace, 9280 dims.
  - (W)PCA: classic (whitened) PCA
  - ML: metric learning first, then regr.
  - JL: proposed, project to 32 dims.

- Conclusion
  - PCA subspaces not effective
  - ML needs more target data
  - JL consistently improves others

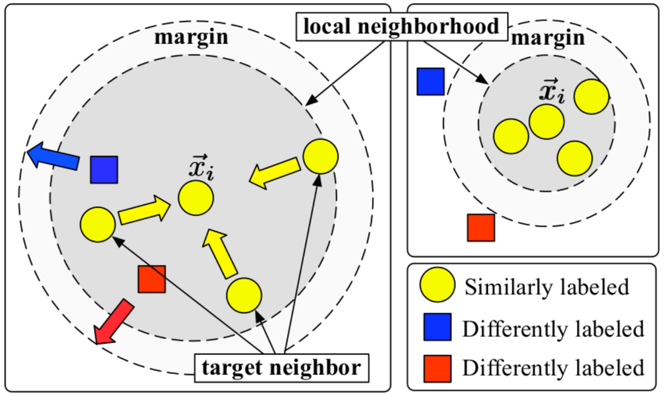
Target Size	Method	Mean of MAE (years)
0	LBP	6.81 ± 0.75
	PCA	7.34 ± 0.73
	WPCA	7.38 ± 0.69
10	LBP	6.82 ± 0.74
	PCA	7.36 ± 0.76
	WPCA	7.40 ± 0.71
	ML	7.20 ± 0.66
	JL	$\mathbf{6.73 \pm 0.73}$
20	LBP	6.69 ± 0.67
	PCA	7.31 ± 0.77
	WPCA	7.35 ± 0.72
	ML	6.66 ± 0.54
	JL	$\mathbf{6.46 \pm 0.62}$
50	LBP	6.46 ± 0.50
	PCA	7.20 ± 0.72
	WPCA	7.25 ± 0.71
	ML	6.21 ± 0.42
	JL	$\mathbf{6.15 \pm 0.44}$

Overview of the presentation

- Face representation
  - Using facial landmarks
  - Aggregated low-level statistics
  - Convolutional networks
  - Comparison



- Metric learning
  - Mahalanobis distances
  - Hierarchical metric learning
  - Local metric learning



- Age estimation

- Conclusion

Conclusion

- Face representations
  - Unsupervised: generic local feature aggregation outperforms landmark based methods
  - Supervised: convolutional neural nets better than unsupervised, amount of training data important

- Metric learning significantly improves performance
  - In particular for unsupervised methods
  - Local metric learning can improve further

- Challenges
  - Dealing with occlusions of parts of the face
  - Matching faces under big pose changes: frontal vs. profile
  - Matching between sketches and photos

References

- B. Bhattarai, G. Sharma, F. Jurie, P. Perez. “Some faces are more equal than others: Hierarchical organization for accurate and efficient large-scale identity-based face retrieval”, ECCV Workshops, 2014.
- G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray. “Visual categorization with bags of keypoints”, ECCV Workshops, 2004.
- N. Dalal, B. Triggs. “Histograms of Oriented Gradients for Human Detection”, CVPR, 2005.
- M. Everingham, J. Sivic, A. Zisserman. “Hello! My name is... Buffy” - automatic naming of characters in TV video”, BMVC, 2006.
- P. Felzenszwalb, D. Huttenlocher. “Pictorial Structures for Object Recognition”, IJCV, 2005.
- R. Girshick and J. Donahue and T. Darrell and J. Malik. “Rich feature hierarchies for accurate object detection and semantic segmentation”, CVPR, 2014.
- M. Guillaumin, J. Verbeek, C. Schmid. “Is that you? Metric learning approaches for face identification”, ICCV, 2009.
- H. Jegou, C. Schmid, H. Harzallah, J. Verbeek. “Accurate image search using the contextual dissimilarity measure”, PAMI 2010.
- A. Krizhevsky, I. Sutskever, G. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”, NIPS 2012.
- Y. LeCun and B. Boser and J. Denker and D. Henderson and R. Howard and W. Hubbard and L. Jackel. “Handwritten Digit Recognition with a Back-Propagation Network”, NIPS 1989.

References

- D. Lowe. “Distinctive image features from scale-invariant keypoints.”, IJCV, 2004.
- J. Sivic, A. Zisserman, “Video Google: a Text Retrieval Approach to Object Matching in Videos”, ICCV, 2003.
- Y. Sun, X. Wang, X. Tang. “Deep Learning Face Representation by Joint Identification-Verification”, CVPR, 2014.
- O. Parkhi and A. Vedaldi and A. Zisserman. “Deep face recognition”, BMVC, 2015.
- F. Perronnin, J. Sánchez, T. Mensink. “Improving the Fisher Kernel for Large-Scale Image Classification”, ECCV, 2010.
- F. Schroff, D. Kalenichenko, J. Philbin. “FaceNet: A Unified Embedding for Face Recognition and Clustering”, CVPR, 2015
- G. Sharma, S. Hussain, F. Jurie. “Local Higher-Order Statistics (LHS) for Texture Categorization and Facial Analysis”, ECCV, 2012.
- K. Simonyan, O. Parkhi, A. Vedaldi, A. Zisserman “Fisher vector faces in the wild”, BMVC, 2013.
- Y. Taigman and M. Yang and M. Ranzato and L. Wolf. “DeepFace: Closing the Gap to Human-Level Performance in Face Verification”, CVPR 2014.
- K. Weinberger, J. Blitzer, L. Saul. “Distance Metric Learning for Large Margin Nearest Neighbor Classification”, NIPS, 2006.