Computer Vision Project – Summer Term 2021

Face Recognition

Dr.-Ing. Thomas Köhler
Pattern Recognition Lab, Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg
Augmented Reality & Cognitive Services, e.solutions GmbH, Erlangen
May 31, 2021

Face Recognition Use Cases

Microsoft Windows Hello: Windows login via face recognition[1]

[1] https://www.microsoft.com/de-de/windows/windows-hello

Face Recognition Use Cases

Apple Face ID: unlock smart phone via face recognition[2]

[2] https://www.apple.com/iphone-x/#face-id

Face Recognition Use Cases

Identification of Vehicle Occupants and Personalization of Vehicle Settings

Scope of this Project

In the lecture:

- Learn how to design and evaluate the core of current facial recognition systems from a technical point of view
- Overview of modern machine learning methods in this field
- Discussion of strengths and limitations of such methods

In the exercise:

- Implementation of a simple system comprising basic functionality for face verification, identification, and clustering from webcam videos
- Evaluation of face recognition algorithms

Outline

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

FACULTY OF ENGINEERING

Introduction

The Face as a Biometric Marker

Biometric markers for identification[3]:

- Fingerprint
- Iris
- Speech
- Face

Face vs. other markers:

- Face the only marker that can be captured at large distances
- Only low-cost hardware required

[3] Jain, Anil K., and Stan Z. Li. Handbook of face recognition. New York, Springer, 2011.

Challenges

Intrinsic conditions:

- High intra-subject variation: facial expressions, eye classes, changes in facial hair, changes in pose, aging
- Low inter-subject variation: similar skin or hair color, similar eye glasses

Extrinsic conditions:

- Varying illumination conditions: images captured at day and at night
- Image quality: images/videos processed with different codecs

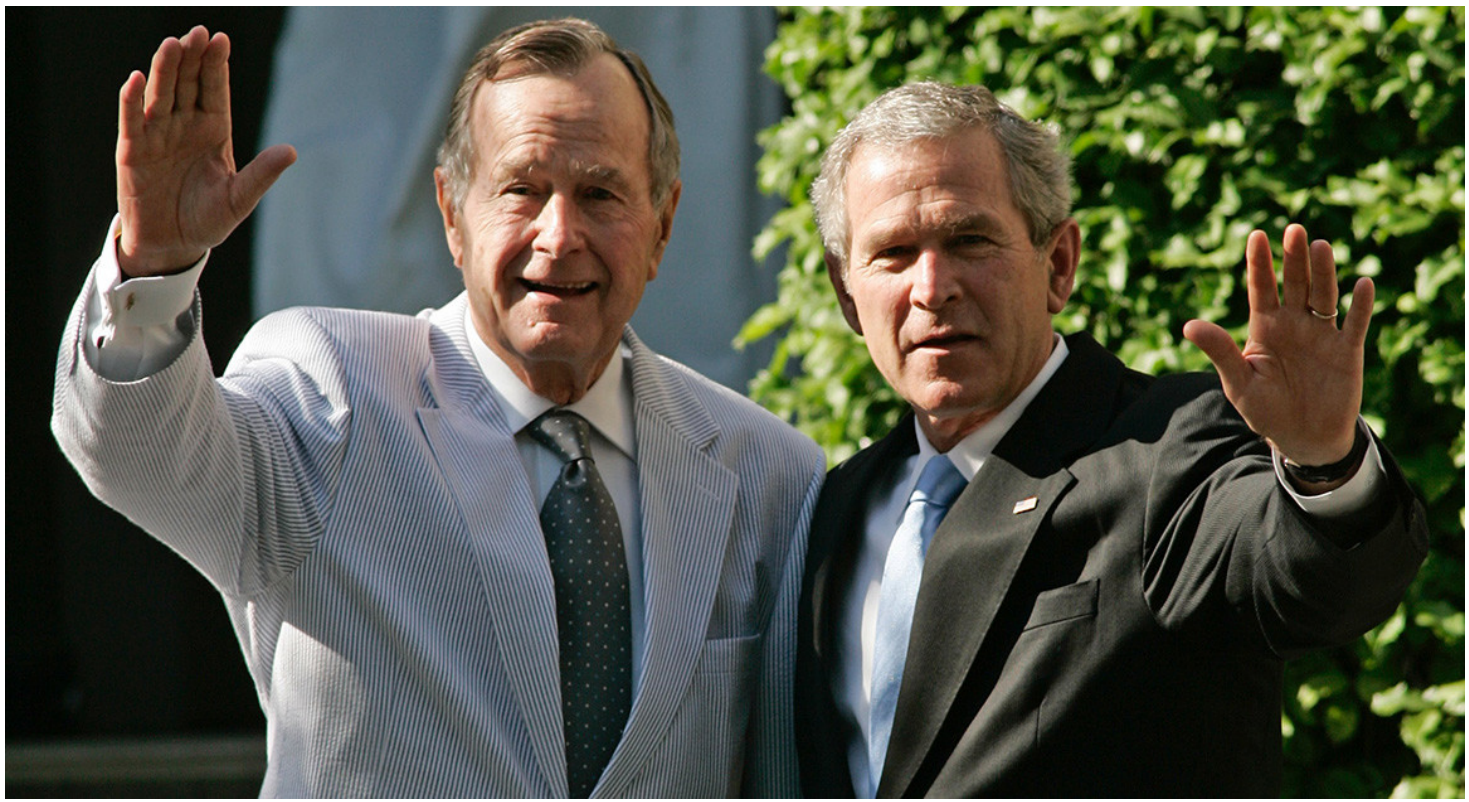
Some Difficult Examples

Frontal and profile views[4]:

[4] Example from CMU Multi-PIE database: http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html

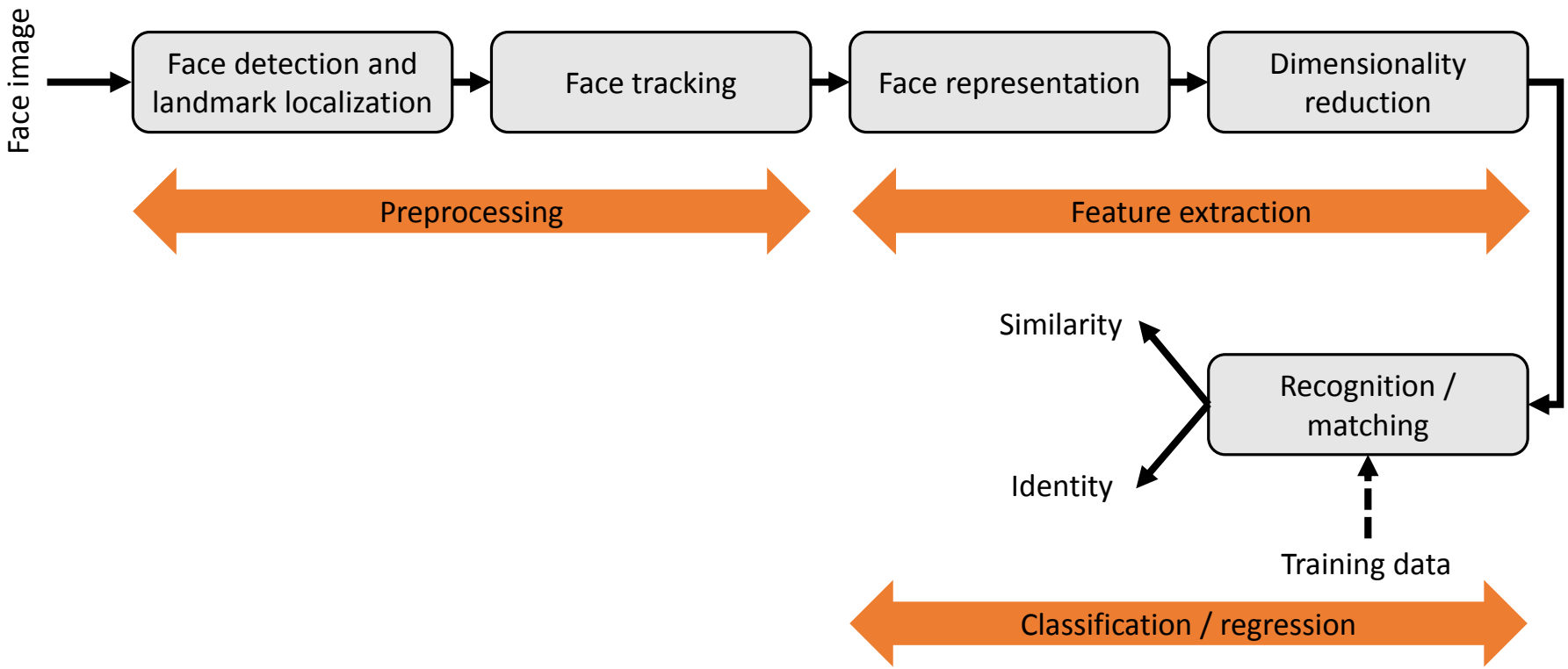
Some Difficult Examples

Close family relationships (e. g. father and son, twins):



Face Recognition Pipeline

Approaching face recognition via machine learning techniques:

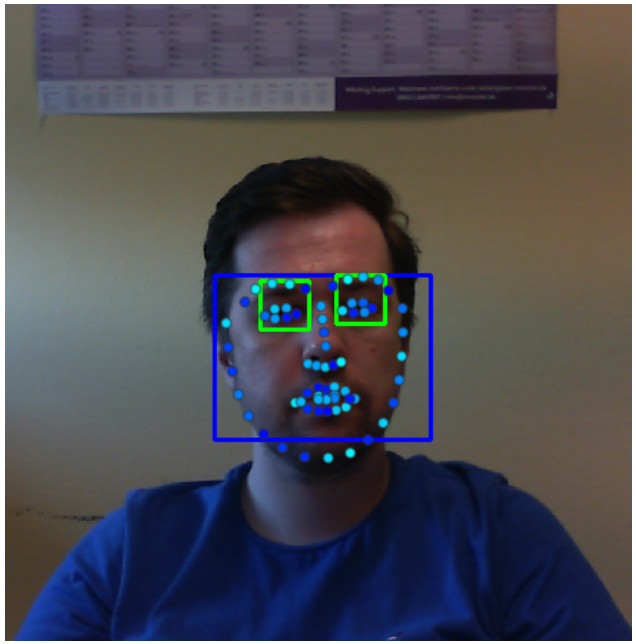


Adopting common pattern recognition pipeline: preprocessing, feature extraction, classification/regression

Face Detection and Facial Landmark Localization

Initial steps of the face recognition pipeline:

- Detection of face region
- Detection of eye regions
- Extracting landmarks to model pose and facial expression
- Alignment of face image using detected landmarks for data normalization
  - Compensate for different head poses
  - Eyes, nose, and mouth at predefined positions



Face Tracking

Continuous monitoring of a human face over time:

- Track face bounding box and landmark positions over time in video data
- 2-D approach: template matching and related motion estimation techniques
- 3-D approach: constrained local model (cf. active shape models (AAM))
  - Describes shape variations of the face
  - Learned from annotated face exemplars
  - Enables the estimation of the head pose

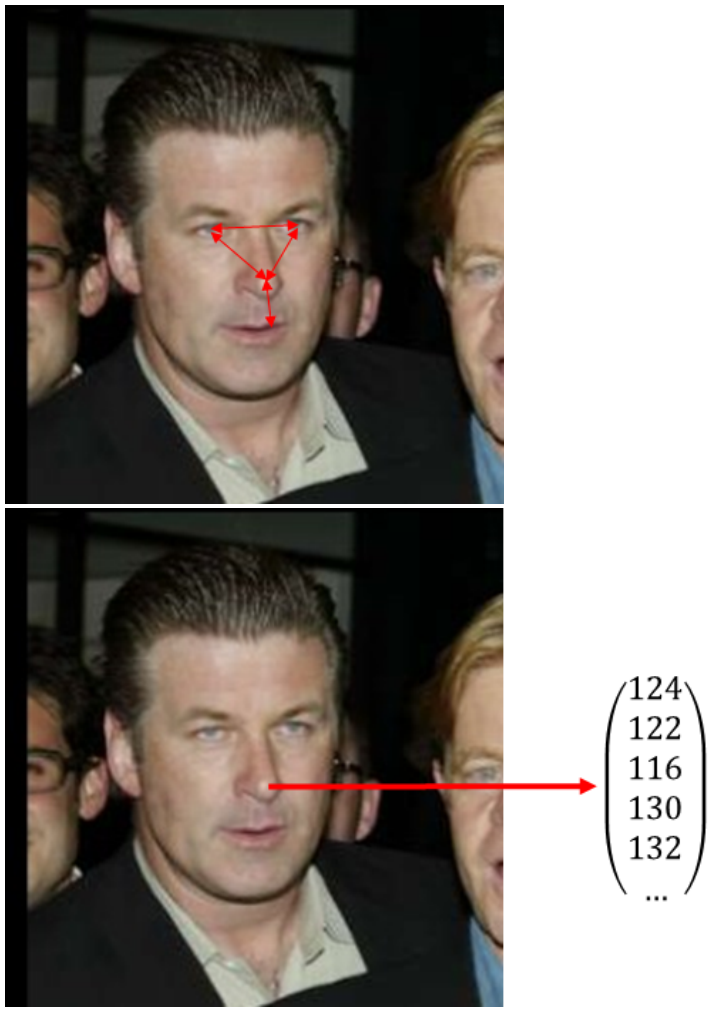
Methodologies of Face Recognition Systems

Geometric face representations:

- Distances, areas, or angles between salient points (eyes, nose, mouth)
- Obtained by feature detection algorithms
- Requires engineering of meaningful features (hand-crafted features)

Data-driven face representations:

- Image intensities form raw features
- Learn suitable representation from exemplars (face manifold)
- Requires training from large datasets



Problem Statements in Face Recognition

Verification (one-to-one matching):

- Given one probe image and one gallery image
- Check if both image show the same identity

Identification (one-to-many matching):

- Given one probe image and a set of gallery images with identity labels
- Retrieve identity of probe image

Clustering (many-to-many matching):

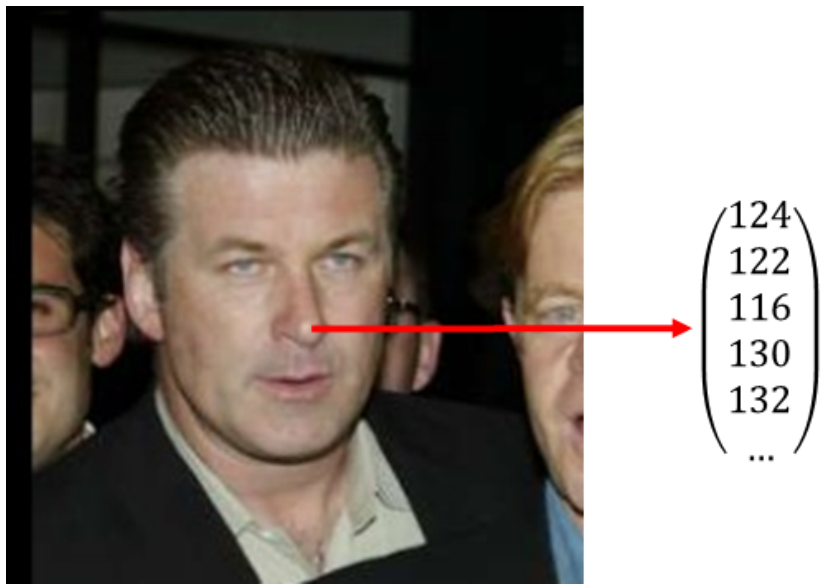
- Given a set of unlabeled face images
- Cluster all images according to the identities captured in the data

FACULTY OF ENGINEERING

Face Representation

Data-Driven Face Representations

How to represent faces in digital images?

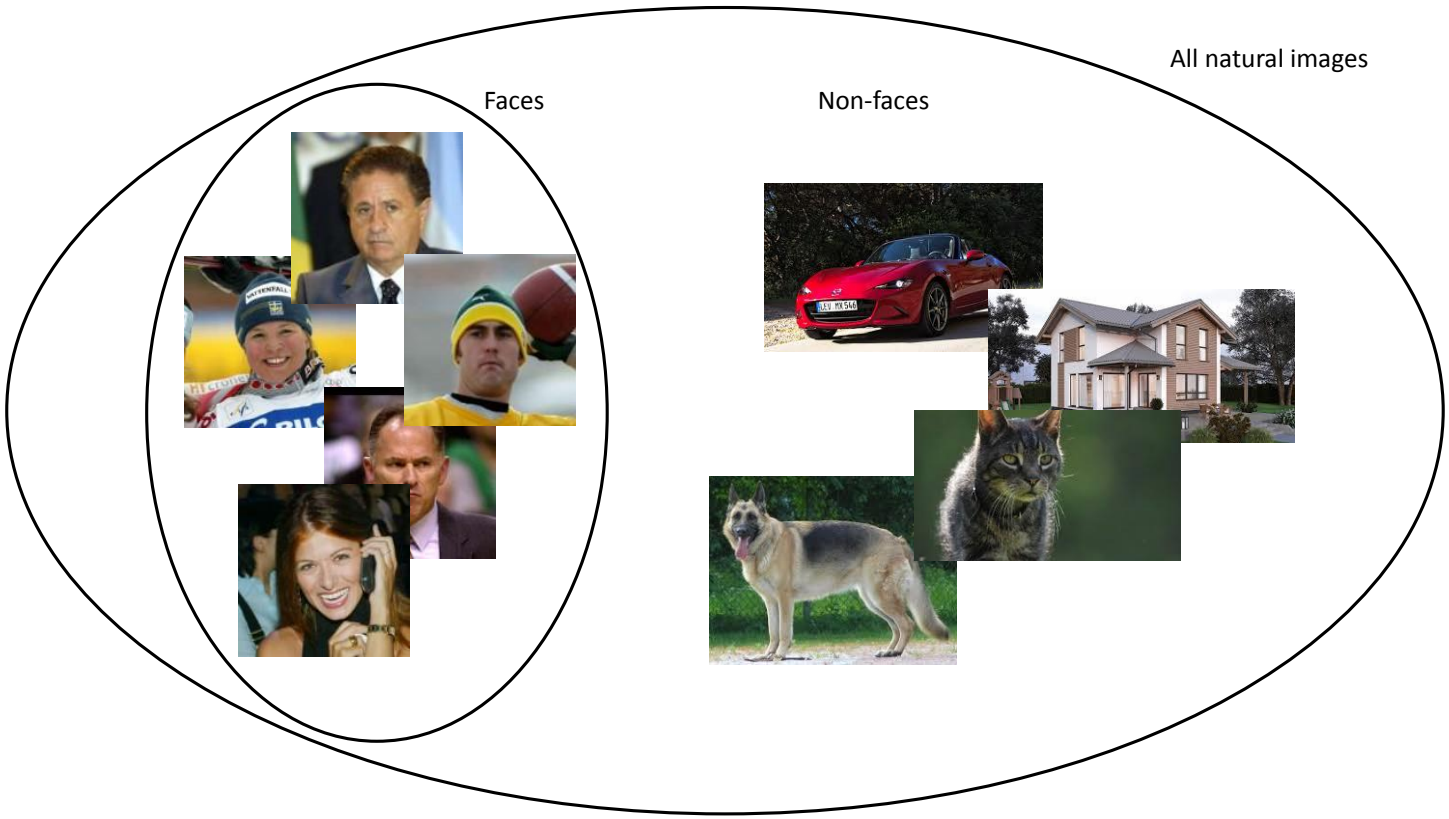


- Graycale image with $M \times N$ pixels and b bits per pixel (e. g. 64×64, 8 bit) $\rightarrow (2^b)^{M \cdot N}$ different images (e. g. $256^{4096} \gg 10^{1000}$)
- Only a small fraction of images corresponds to valid faces

Manifold of Face Images

Faces live on a low-dimensional manifold

- Learn the manifold form exemplar data
- Face detection: distinguish faces and non-faces
- Face recognition: distinguish faces of different subjects



Eigenfaces

- Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, $\boldsymbol{x}_i \in \mathbb{R}^D$ be a set of n face images with mean:

$$\mu_x = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \tag{1}$$

  and covariance:

$$\Sigma_x = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \mu_x)(\boldsymbol{x}_i - \mu_x)^\top \tag{2}$$

- Project each $\boldsymbol{x}_i$ to a M-dimensional space ($M \ll D$):

$$\boldsymbol{y}_i = \boldsymbol{W}\boldsymbol{x}_i \in \mathbb{R}^M \tag{3}$$

- Seek linear transform $\boldsymbol{W} \in \mathbb{R}^{M \times D}$ that maximizes the variance of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$:

$$\begin{aligned} \Sigma_y &= \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{y}_i - \mu_y)(\boldsymbol{y}_i - \mu_y)^\top \\ &= \boldsymbol{W}^\top \Sigma_x \boldsymbol{W} \end{aligned} \tag{4}$$

  $\rightarrow$ Principal Component Analysis (Principal components $\equiv$ Eigenfaces)

Fisherfaces

- Eigenfaces ignore class labels to construct the feature transform
- Fisherfaces exploit two constraints (one unique face $\equiv$ one class):
  1. The between-class scatter is maximized
  2. The within-class scatter is minimized
- Seek the transform $\boldsymbol{W} \in \mathbb{R}^{M \times D}$ that maximizes Fisher's linear discriminant:

$$J(\boldsymbol{W}) = \frac{\boldsymbol{W}^\top \Sigma_{\text{inter}} \boldsymbol{W}}{\boldsymbol{W}^\top \Sigma_{\text{intra}} \boldsymbol{W}} \tag{5}$$

Intra-class and inter-class scatter for c classes:

$$\Sigma_{\text{intra}} = \sum_{i=1}^{c} \frac{1}{|\mathscr{X}_i|} \sum_{\boldsymbol{x}_k \in \mathscr{X}_i} (\boldsymbol{x}_k - \mu_i)(\boldsymbol{x}_k - \mu_i)^\top \tag{6}$$

$$\Sigma_{\text{inter}} = \sum_{i=1}^{c} (\mu_i - \mu)(\mu_i - \mu)^\top \tag{7}$$

μ_i is the mean of all faces $\mathscr{X}_i$ in the i-th class and μ is the overall mean face

Deep Features

Limitations of the methods discussed so far:

- Determine face representations under a linear model
- Linear models feature limited robustness against pose or illumination variations

Extension:

- Non-linear face representation $f_\theta(\boldsymbol{x})$:

$$\boldsymbol{y} = f_\theta(\boldsymbol{x}) \tag{8}$$

- The transform $f_\theta(\boldsymbol{x})$ is implemented by a deep neural network
- Parameters θ are learned from exemplars

Convolutional Neural Networks (CNNs)

Neural network:

- Computation graphs comprising neurons
- Propagation of input signals through the network to obtain outputs

Network design with input, output, and hidden layers:

- Convolutional layer: convolution of input neuron activations with filter kernel
- Pooling layer: fusion of clusters of input neuron activations
- Locally/fully connected layer: weighted sum of all input neuron activations

θ: Parameters of all layers in the network

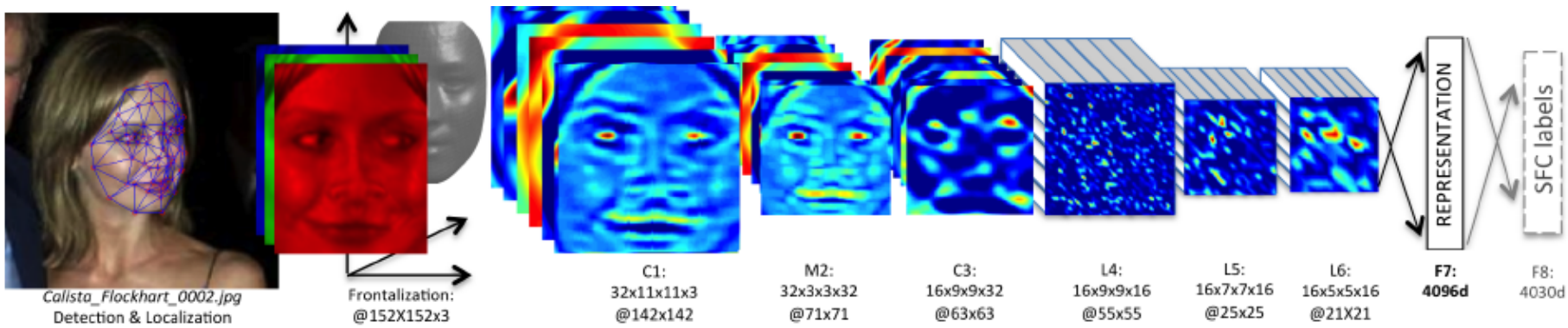
Face Representation via Classification

- Design neural network to classify face images according to their identity
- Fully connected output layer $\left(\boldsymbol{W}, \boldsymbol{b}\right)$ to model probability distribution over c classes:

$$p_i = \frac{\exp\left(\boldsymbol{w}_i^\top \boldsymbol{x} + b_i\right)}{\sum_{j=1}^{c} \exp\left(\boldsymbol{w}_j^\top \boldsymbol{x} + b_j\right)}, \qquad i = 1, \ldots, c \tag{9}$$

  $\rightarrow$ Softmax activation
- Train network parameters on face exemplars by minimizing misclassification error (e.g. cross entropy loss)
- Activation of fully connected layer $\equiv$ face representation (aka. embedding)

Example: Facebooks DeepFace (trained on 4M faces)[5]



- Face frontalization for preprocessing
- Eight layer network: convolutional (C1, C3), max-pooling (M2), locally connected (L4, L4, L6), and fully connected layer (F7, F8)
- Feature maps model face from high-level to low-level
- Activation of F7 fully connected layer used as face representation

[5] Taigman, Yaniv, et al. "Deepface: Closing the gap to human-level performance in face verification." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014.

Face Representation via Regression

Discussion of the classification-based approach:

- Classification requires class labels (identities)
- Models discriminative features for face recognition only implicitly
- Learned features are not necessarily optimal

Extension to regression-based face representation:

- Training of deep neural networks for classification
- Fine-tuning by regression-based loss

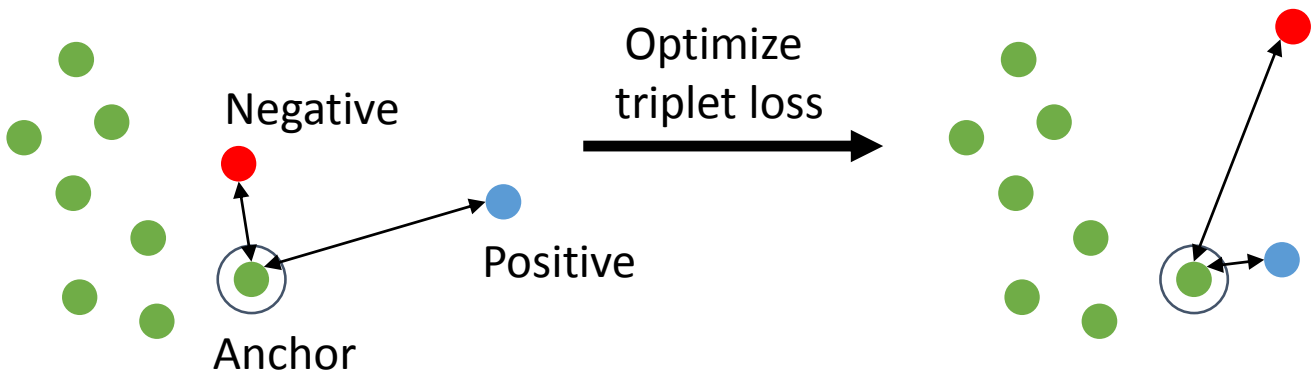
Triplet Loss

- Let $\boldsymbol{x}_i^a$ be an anchor face, $\boldsymbol{x}_i^p$ a face of the same subject (positive) and $\boldsymbol{x}_i^n$ be a face of different subject (negative)
- Ensure for all triplets $\left(\boldsymbol{x}_i^a, \boldsymbol{x}_i^p, \boldsymbol{x}_i^n\right)$ with margin α:

$$\underbrace{||f_\theta(\boldsymbol{x}_i^a) - f_\theta(\boldsymbol{x}_i^p)||_2^2}_{\text{anchor-to-positive distance}} + \alpha < \underbrace{||f_\theta(\boldsymbol{x}_i^a) - f_\theta(\boldsymbol{x}_i^n)||_2^2}_{\text{anchor-to-negative distance}} \tag{10}$$

- Penalize distances of positive and negative samples w.r.t. the anchor

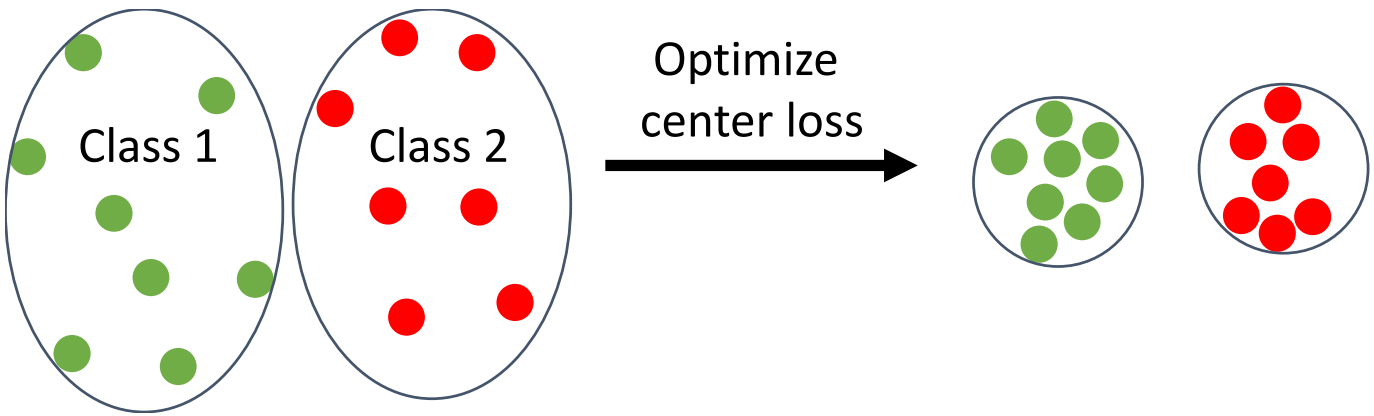
$$\mathscr{L}_{\text{triplet}}(\theta) = \sum_{i=1}^{n} ||f_\theta(\boldsymbol{x}_i^a) - f_\theta(\boldsymbol{x}_i^p)||_2^2 - ||f_\theta(\boldsymbol{x}_i^a) - f_\theta(\boldsymbol{x}_i^n)||_2^2 + \alpha \tag{11}$$



Center Loss

- Penalize variations of deep features $\boldsymbol{x}_i$ around their class center (mean) μ_i

$$\mathscr{L}_{\text{center}}(\theta) = \sum_{i=1}^{n} ||f_\theta(\boldsymbol{x}_i) - \mu_i||_2^2 \tag{12}$$



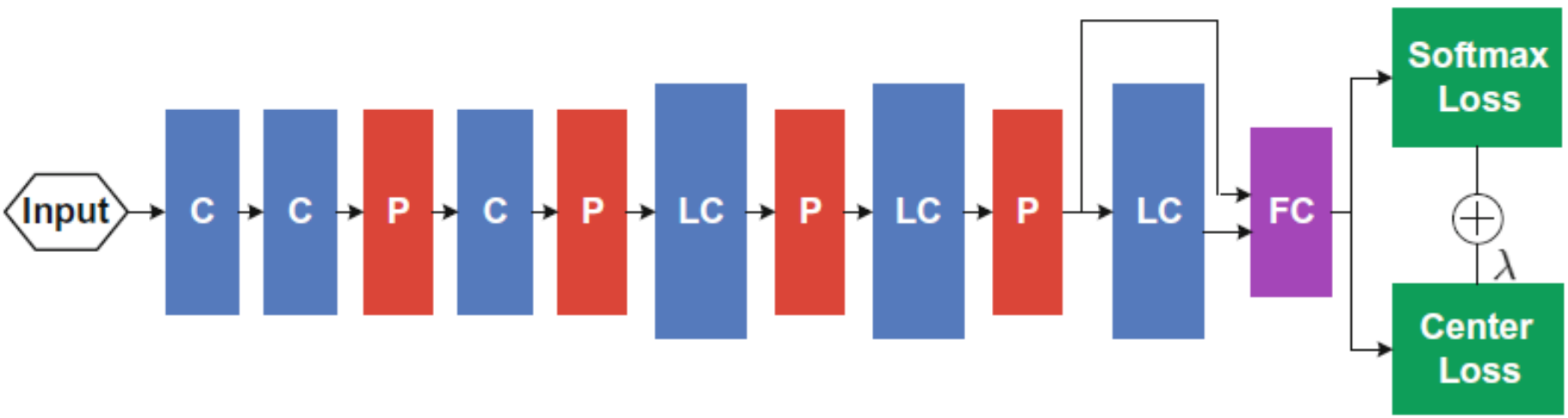
- Similar to Fisherfaces but used with a non-linear model
- Does not require the handling of triplets (combinatorial explosion)

Combining Classification and Regression Approaches

Joint supervision with cross entropy and center loss:

$$\mathscr{L}(\theta) = \mathscr{L}_{\text{ce}}(\theta) + \lambda \mathscr{L}_{\text{center}}(\theta) \tag{13}$$

Example architecture of Wen et al.[6]:

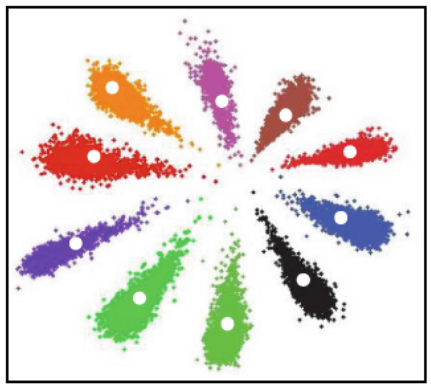


Network with convolutional (C), max-pooling (P), local connected (LC), and fully connected layer (FC)

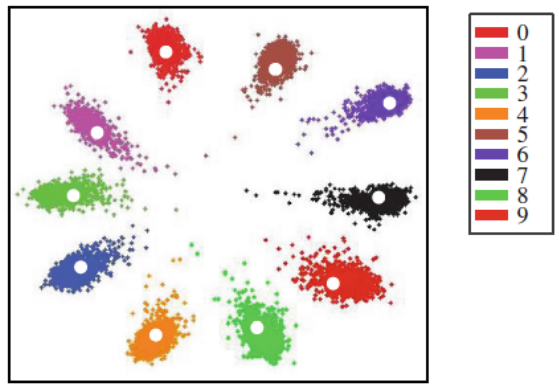
[6] Wen, Yandong, et al. "A discriminative feature learning approach for deep face recognition." European Conference on Computer Vision. Springer, 2016.

Combining Classification and Regression Approaches

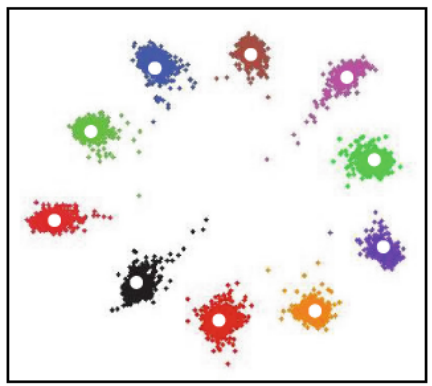
Controlling discriminative power of deep features in the method of Wen et al.:



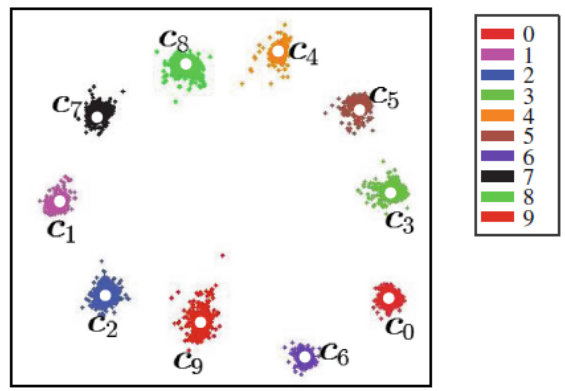
(a) $\lambda = 0.001$



(b) $\lambda = 0.01$



(c) $\lambda = 0.1$



(d) $\lambda = 1$

Larger center loss weight (large λ) $\Rightarrow$ higher discriminative power of deep features

Practical Considerations on Deep Features

Learning:

- Learn $f_\theta(\boldsymbol{x})$ on large datasets (Facebook: 4M faces, Google: 200M faces)
- Computationally very demanding
- Implemented on graphics processing units (GPU)

Inference:

- Inference of face representation $f_\theta(\boldsymbol{x})$ by forward pass
- Efficient to compute using additional optimizations (weight quantization)
- Use face representation within lightweight classification or clustering models
- Can be implemented on embedded devices

FACULTY OF ENGINEERING

Selected Topics in Face Recognition

Face Verification

- Given two face images $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, is the image pair of the same person?
- Define a suitable distance measure $d(\boldsymbol{x}_1, \boldsymbol{x}_2)$
  - Use face representation instead of the image space
  - Example: cosine distance

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{||f_\theta(\boldsymbol{x}_1) - f_\theta(\boldsymbol{x}_2)||_2^2}{||f_\theta(\boldsymbol{x}_1))||_2^2 + ||f_\theta(\boldsymbol{x}_2)||_2^2} \tag{14}$$

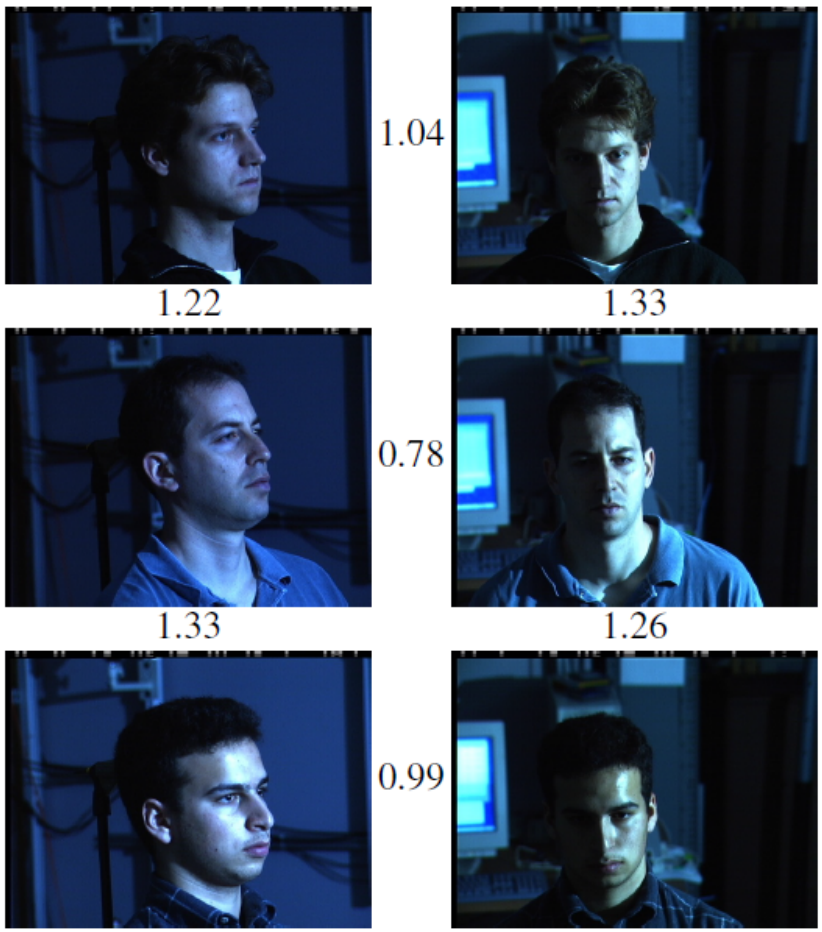
- Verification using the distance measure and threshold $\tau > 0$:

$$v(\boldsymbol{x}_1, \boldsymbol{x}_2) = \begin{cases} \text{same} & \text{if } d(\boldsymbol{x}_1, \boldsymbol{x}_2) \leq \tau \\ \text{not same} & \text{if } d(\boldsymbol{x}_1, \boldsymbol{x}_2) > \tau \end{cases} \tag{15}$$

- Alternatively, use similarity measure, e. g. $s(\boldsymbol{x}_1, \boldsymbol{x}_2) = -d(\boldsymbol{x}_1, \boldsymbol{x}_2)$

Face Verification

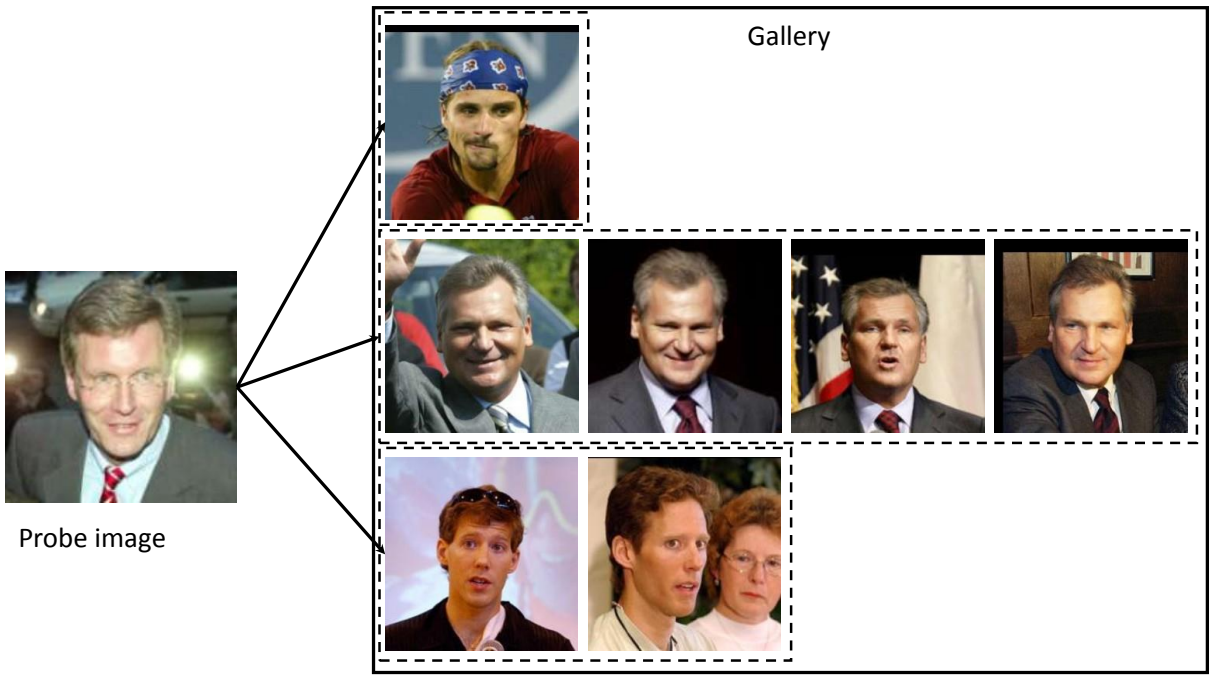
Example: Face identification using Google FaceNet embeddings[7]



[7] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

Face Identification

- Given a probe image $\boldsymbol{x}_{\text{probe}}$ of unknown identity
- Determine identity $I(\boldsymbol{x}_{\text{probe}})$ from labeled images in a gallery



Two different protocols:

- Closed-set: all possible identities of probes are contained in the gallery
- Open-set: identities of some probes are missing in the gallery

Face Identification with Closed-Set Protocol

Repeated face verification using k-nearest neighbors (k-NN) classifier:

- For the probe image $\boldsymbol{x}_{\text{probe}}$, find the k closest gallery images $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ according to a distance $d(\boldsymbol{x}_{\text{probe}}, \cdot)$
- Identity of $\boldsymbol{x}_{\text{probe}}$ is the identity of the majority (mode) in $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$:

$$I(\boldsymbol{x}_{\text{probe}}) = \text{mode}(I(\boldsymbol{x}_1), \ldots, I(\boldsymbol{x}_k)) \tag{16}$$

- Alternatively, we can consider similarities $s(\boldsymbol{x}_{\text{probe}}, \cdot) = -d(\boldsymbol{x}_{\text{probe}}, \cdot)$

Other discriminative classification models:

- Support vector machine (SVM)
- Random forests
- Boosting methods

Face Identification with Open-Set Protocol

Handle open-set space with distance-based approach (thresholded NN):

- In addition, extract the best matching gallery image $\boldsymbol{x}_{\text{match}}$ with minimum distance to $\boldsymbol{x}_{\text{probe}}$
- Possibly assign "unknown" identity to $\boldsymbol{x}_{\text{probe}}$:

$$I(\boldsymbol{x}_{\text{probe}}) = \begin{cases} \text{mode}\big(I(\boldsymbol{x}_1), \ldots, I(\boldsymbol{x}_k)\big) & \text{if } d(\boldsymbol{x}_{\text{probe}}, \boldsymbol{x}_{\text{match}}) \leq \tau \\ \text{unknwon} & \text{if } d(\boldsymbol{x}_{\text{probe}}, \boldsymbol{x}_{\text{match}}) > \tau \end{cases} \tag{17}$$

- τ is the face verification threshold

Handle open-set space via extreme value theory[8]:

- Statistical model for inclusion probability of probe images in classes contained in the gallery
- Thresholding of inclusion probabilities instead of using distances

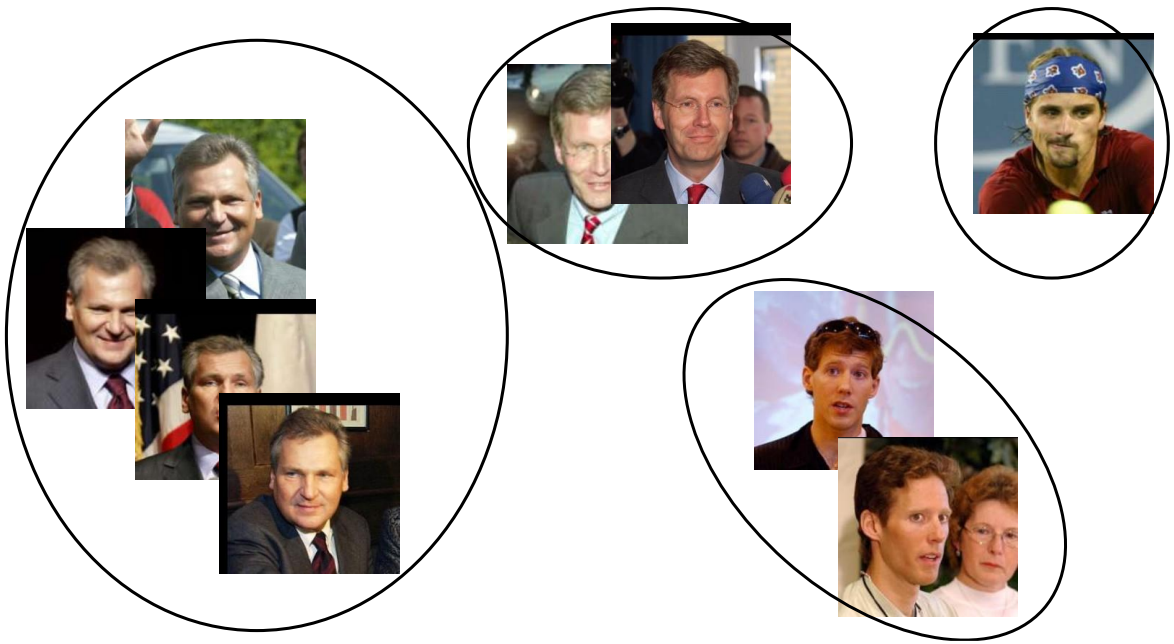
[8] Günther, Manuel, et al. "Toward open-set face recognition." Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2017.

Face Clustering

- Given n face images $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, cluster them into $k \leq n$ clusters $\mathscr{C}_1, \ldots, \mathscr{C}_k$
- Minimize cluster scatter with cluster centers μ_i and distance measure $d(\cdot, \cdot)$:

$$(\mathscr{C}_1, \ldots, \mathscr{C}_k) = \text{argmin}_{\mathscr{C}_1,\ldots,\mathscr{C}_k} \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in \mathscr{C}_i} d(\boldsymbol{x}, \mu_i) \tag{18}$$

- k-means clustering: $d(\boldsymbol{x}, \mu_i) = ||\boldsymbol{x} - \mu_i||_2^2$ and $\mu_i \equiv$ cluster mean



Clustering using k-Means Algorithm

Iterative algorithm:

1. Assignment: Assign each face to the cluster with closest center

$$\mathscr{C}_i^t = \left\{ \boldsymbol{x} : ||\boldsymbol{x} - \boldsymbol{\mu}_i||_2^2 \leq ||\boldsymbol{x} - \boldsymbol{\mu}_j||_2^2 \text{ for all } i \neq j \right\} \tag{19}$$

2. Update: Re-calculate the cluster centers from current clustering $\mathscr{C}_1^t, \ldots, \mathscr{C}_k^t$

$$\boldsymbol{\mu}_i^{t+1} = \frac{1}{|\mathscr{C}_i^t|} \sum_{\boldsymbol{x} \in \mathscr{C}_i^t} \boldsymbol{x} \tag{20}$$

Other clustering approaches[9]:

- Soft clustering: membership degrees instead of assignment to single clusters
- Agglomerative clustering: adaptive selection of the number of clusters

[9] Otto, C., Wang, D., and Jain, A. K. (2018). Clustering millions of faces by identity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(2), 289-303.

Evaluating Face Recognition Systems

Face identification with closed-set protocol:

- Rank of a probe image $\boldsymbol{x}_{\text{probe}}$ for gallery $\mathcal{G}$ with true matching image $\boldsymbol{x}_{\text{match}}$:

$$\text{Rank}(\boldsymbol{x}_{\text{probe}}) = \left| \left\{ \boldsymbol{x}' \in \mathcal{G} \ : \ s(\boldsymbol{x}', \boldsymbol{x}_{\text{probe}}) \geq s(\boldsymbol{x}_{\text{match}}, \boldsymbol{x}_{\text{probe}}) \right\} \right| \tag{21}$$

  $\text{Rank}(\boldsymbol{x}_{\text{probe}}) = 1$ if $\boldsymbol{x}_{\text{probe}}$ is correctly associated with $\boldsymbol{x}_{\text{match}}$

- Rank-k Identification rate on a test set $\mathcal{T}$:

$$\text{IR}(r) = \frac{\left| \left\{ \boldsymbol{x}' \in \mathcal{T} \ : \ \text{Rank}(\boldsymbol{x}') \geq r \right\} \right|}{\left| \mathcal{T} \right|} \tag{22}$$

  For $r = 1$ it is equivalent to the accuracy

- Depict the identification rate $\text{IR}(r)$ versus the rank r

Evaluating Face Recognition Systems

Face identification with open-set protocol:

- Trade-off between true identifications and unknowns that are incorrectly detected as knowns (false alarms) depending on similarity threshold θ
- Detection and identification rate for test set of knowns $\mathscr{K}$ at rank r:

$$\mathrm{DIR}(\theta) = \frac{\left|\left\{\boldsymbol{x}' \in \mathscr{K} \;:\; s\left(\boldsymbol{x}', \boldsymbol{x}_{\text{gallery}}\right) \geq \theta \text{ and } \mathrm{Rank}\left(\boldsymbol{x}'\right) \geq r\right\}\right|}{\left|\mathscr{K}\right|} \tag{23}$$

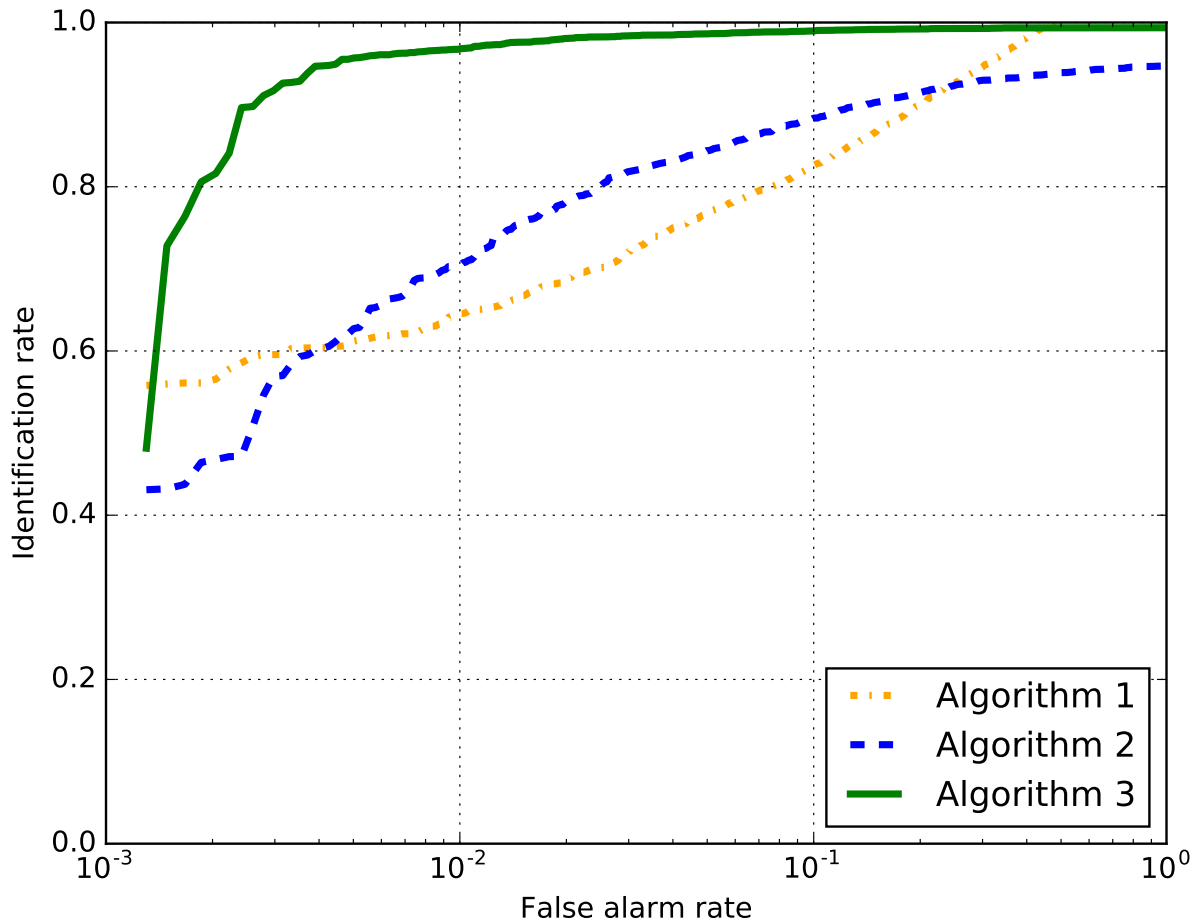
- False alarm rate for complementary test set of unknowns $\mathscr{U}$:

$$\mathrm{FAR}(\theta) = \frac{\left|\left\{\boldsymbol{x}' \in \mathscr{U} \;:\; s\left(\boldsymbol{x}', \boldsymbol{x}_{\text{gallery}}\right) \geq \theta \text{ for any } \boldsymbol{x}_{\text{gallery}} \in \mathscr{G}\right\}\right|}{\left|\mathscr{U}\right|} \tag{24}$$

- Depict $\mathrm{DIR}(\theta)$ at different $\mathrm{FAR}(\theta)$ for a given rank r (typically $r = 1$)

Example: Comparing Different Face Recognition Algorithms

DIR curve with semi-logarithmic axes for three algorithms (shown for rank $r = 1$):



Summary

Take Home Messages

- Widespread application domains of face recognition
- Still a hard problem under uncontrolled conditions (e. g. difficult poses)
- Face representation is a key component for modern face recognition systems
  - Learned on large training datasets
  - Different methodologies: Eigenfaces, Fisherfaces, deep features
- Recognition tasks (verification, identification, clustering) solved by common machine learning algorithms
  - Based on suitable face representation
  - Today also applicable on embedded devices

Further Readings

Overview on face image analysis and recognition techniques:

Anil K. Jain and Stan Z. Li. "Handbook of face recognition". Springer, 2011

Eigenfaces, Fisherfaces, and other classical methods:

Peter N. Belhumeur, João P. Hespanha, and David J. Kriegman. "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection." IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 1997, 711-720.

Deep learning based methods:

- Yaniv Taigman *et al.* "Deepface: Closing the gap to human-level performance in face verification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014
- Florian Schroff, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015
- Yandong Wen *et al.* "A discriminative feature learning approach for deep face recognition." European Conference on Computer Vision, 2016.

Open Project Topics (Collaboration with e.solutions Erlangen)

Improving face representation learning using adversarial samples (5/10 ECTS)

- Adversarial attack: fool face recognition system by manipulating input data
- Project goal: investigate benefit of adversarial samples for data augmentation
- Begin: February 2021

Learning image formation models for super-resolution (5/10 ECTS)

- Learning-based super-resolution aims at inferring mappings from low-resolution to high-resolution images
- Classical reconstruction-based methods use reversed mappings for inverse problem formulations
- Project goal: learn reversed mapping for reconstruction algorithms
- Begin: February 2021

All topics are extendable to a master's thesis

Thanks for listening.

Any questions?