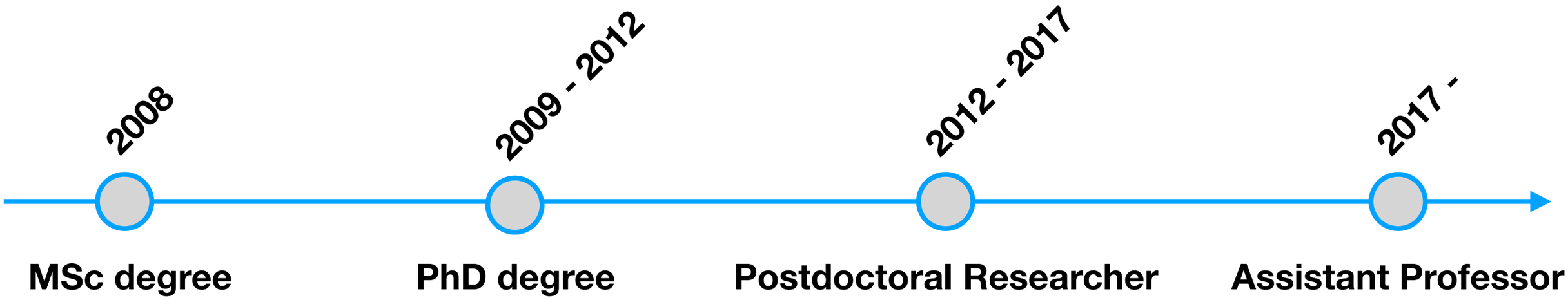
Giuseppe Lisanti

Assistant Professor @ University of Pavia

Bio



MSc degree

Computer Engineering
University of Florence

PhD degree

Computer Engineering,
Multimedia and
Telecommunication
University of Florence

PhD Thesis:
Wide area surveillance with rotating and zooming cameras

Postdoctoral Researcher

Media Integration and
Communication Center
University of Florence

Working on:
Person Re-Identification
3D Face Modeling
Face Recognition

Assistant Professor

Vision Lab - DIII
University of Pavia

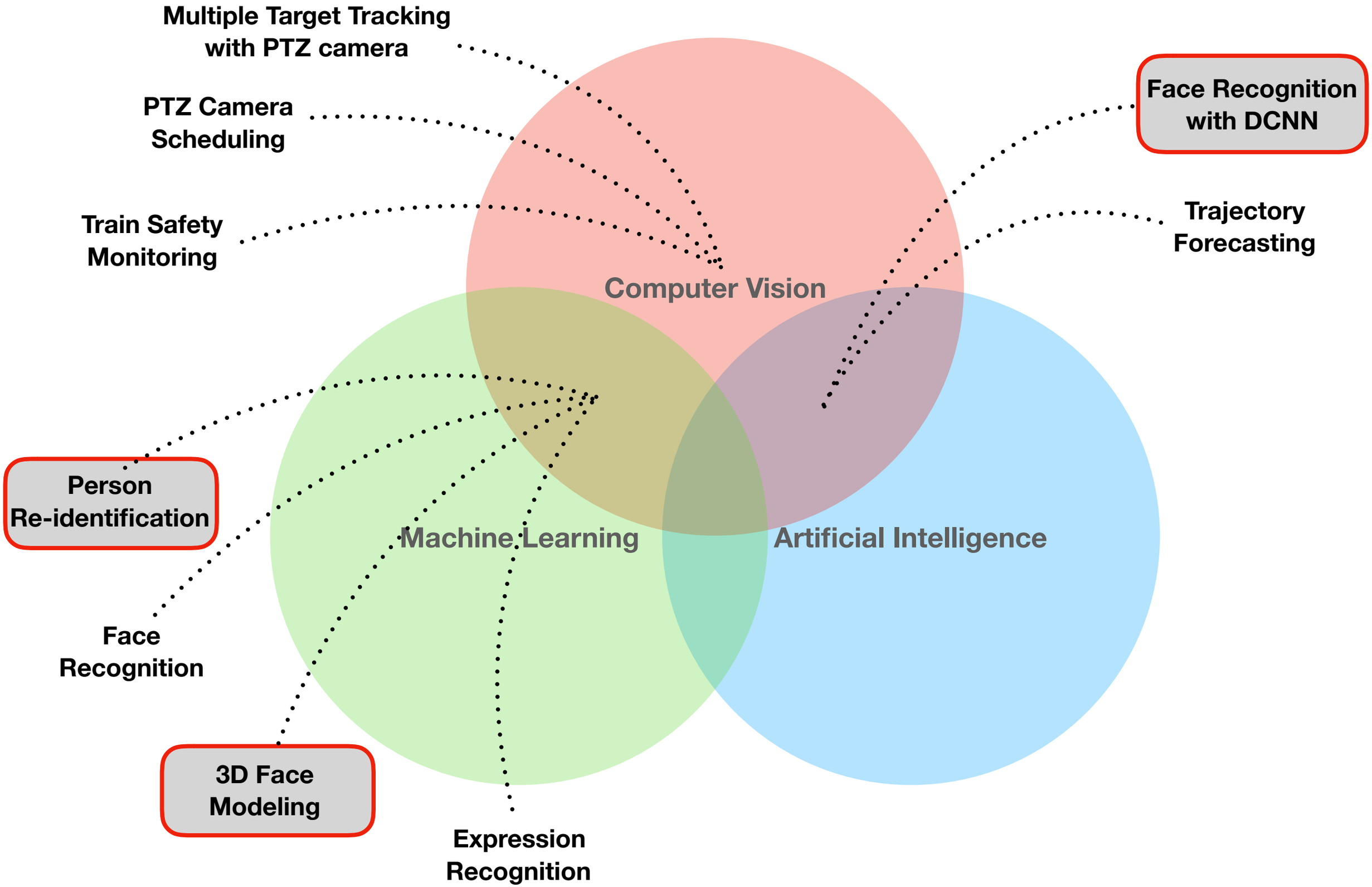
Other Activities

- Collaborations with both national and international research centres

- Several research projects and industry-funded projects:

- International Workshop on Cross-domain Human Identification (ICCV'17)

Research Interests

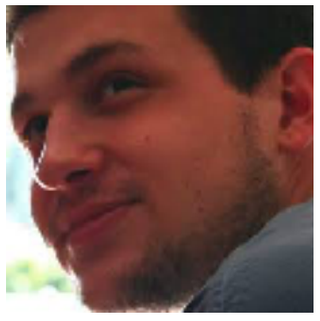


Person Re-Identification

A. D. Bagdanov

I. Masi

S. Karaman



A. Del Bimbo

Person Re-Identification

- Recognizing an individual in diverse locations over different camera views

- **Issues**: people leave the scene for long time, different camera and different viewpoints, occlusions, very low resolution, Illumination variations, pedestrians with very similar clothes

Motivations

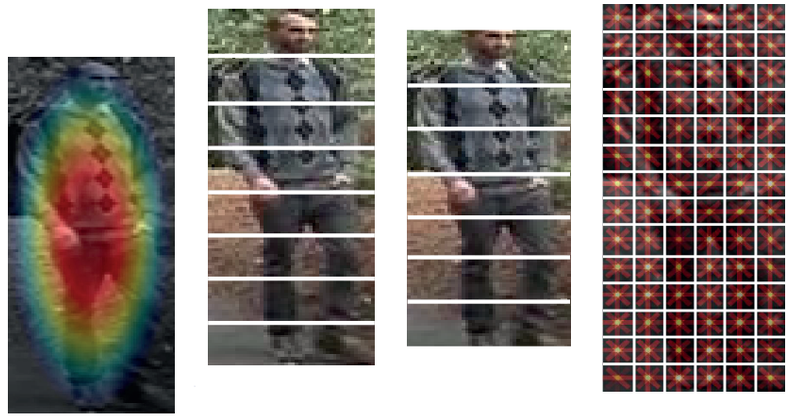
- Can be applied in various real scenarios: airports, train stations, wide area (parking area, museum)

- For the London underground bomb the police **manually** examined about 2,500 items of CCTV footage and forensic evidence from the scenes of the attacks

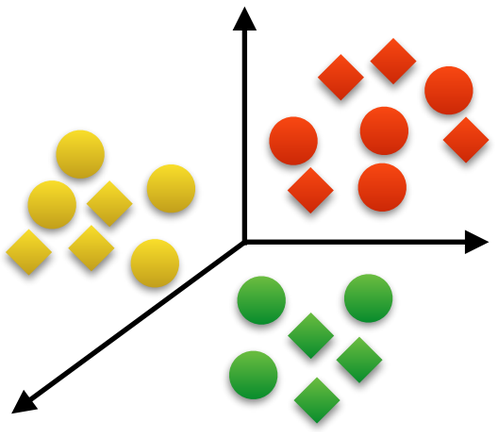
- Smart museum: collect persons locations in a museum and then give additional information to improve user experience, based on its preferences (Mnemosyne Project)

Both appearance and learning are important:

- Compute a descriptor for all the person images, for each camera



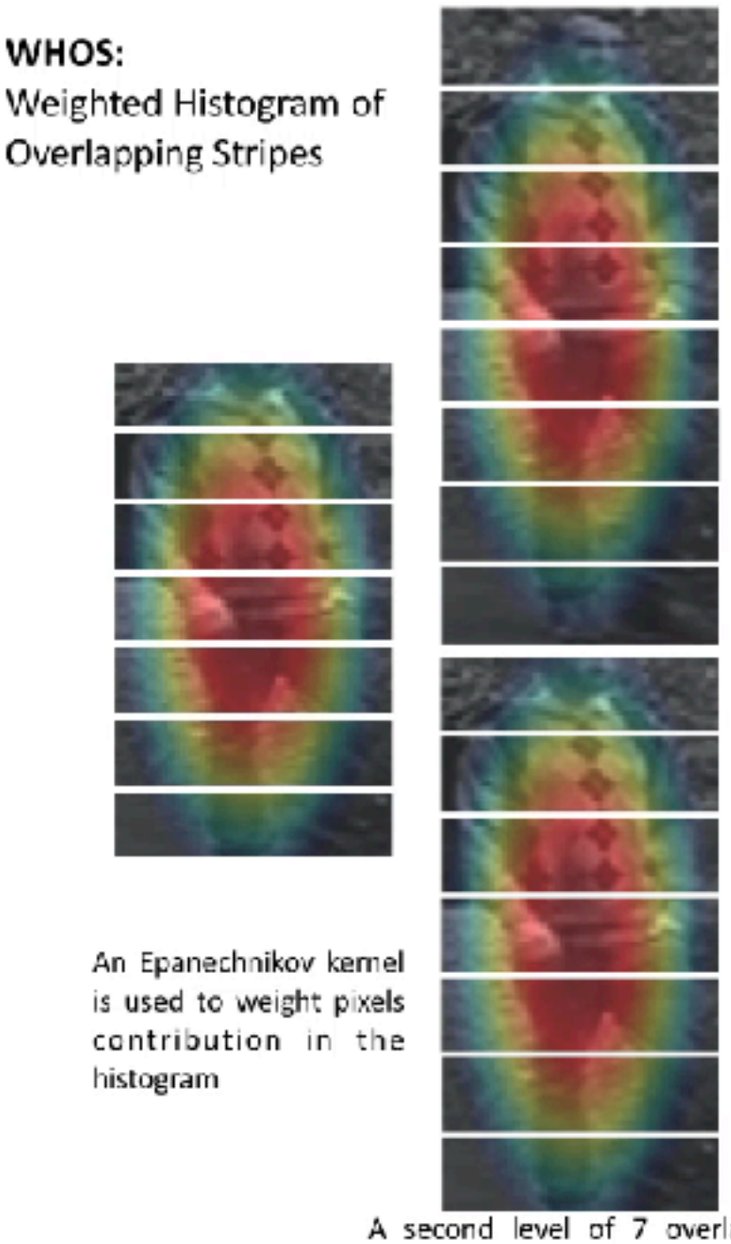
- Apply kernel trick to descriptors and learn a common space between the two views



- Perform person re-identification in the common space

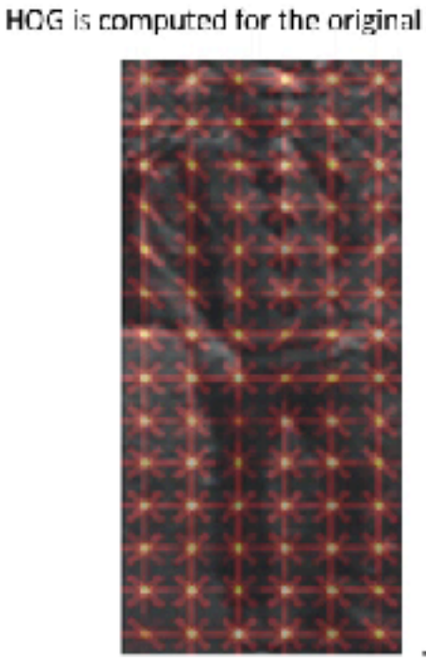
WHOS: Weighted Histogram of Overlapping Stripes

Designed a discriminative and efficient descriptor of person appearance for re-identification based on coarse, striped pooling of local features:



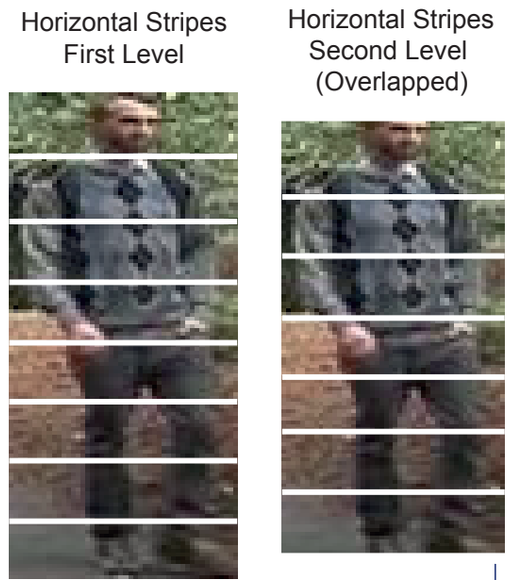
An Epanechnikov kernel is used to weight pixels contribution in the histogram

A second level of 7 overlapped stripes is extracted



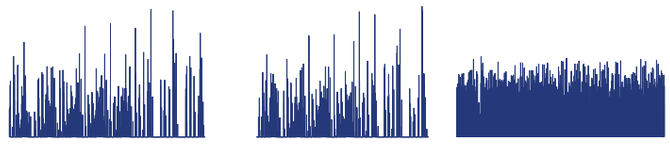
The final descriptor is obtained as the concatenation of the weighted HS and RGB histograms and the HOG descriptor.

WHOS: Weighted Histogram of Overlapping Stripes



- The striped pooling model grants a degree of pose invariance in the person representation:
  - Horizontal stripes capture information about vertical color distribution in the image
  - overlapping stripes maintain color correlation information between adjacent stripes in the final descriptor

- Color information is captured by HS, RGB and Lab histograms, and local texture by the LBP and HOG:
  - The use of HS histograms renders a portion of the descriptor invariant to illumination variations
  - The RGB histograms capture more discriminative color information especially for dark and greyish colors.
  - Gradient histogram over only 4 angular bins (to capture vertical, horizontal and diagonal patterns).

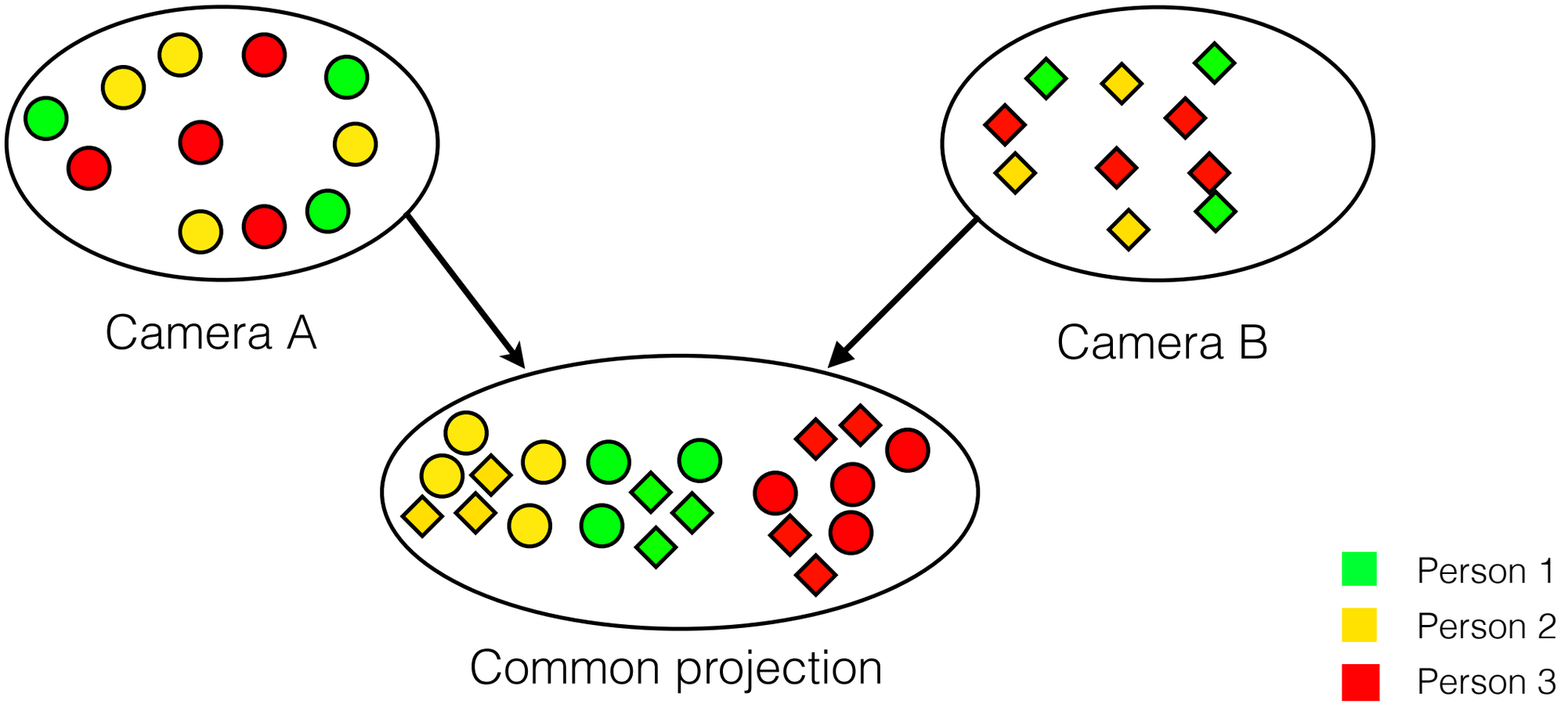


- Pixel contribution to histogram bin is weighted through a non-isotropic Gaussian kernel to decrease background pixel influence.

Canonical Correlation Analysis

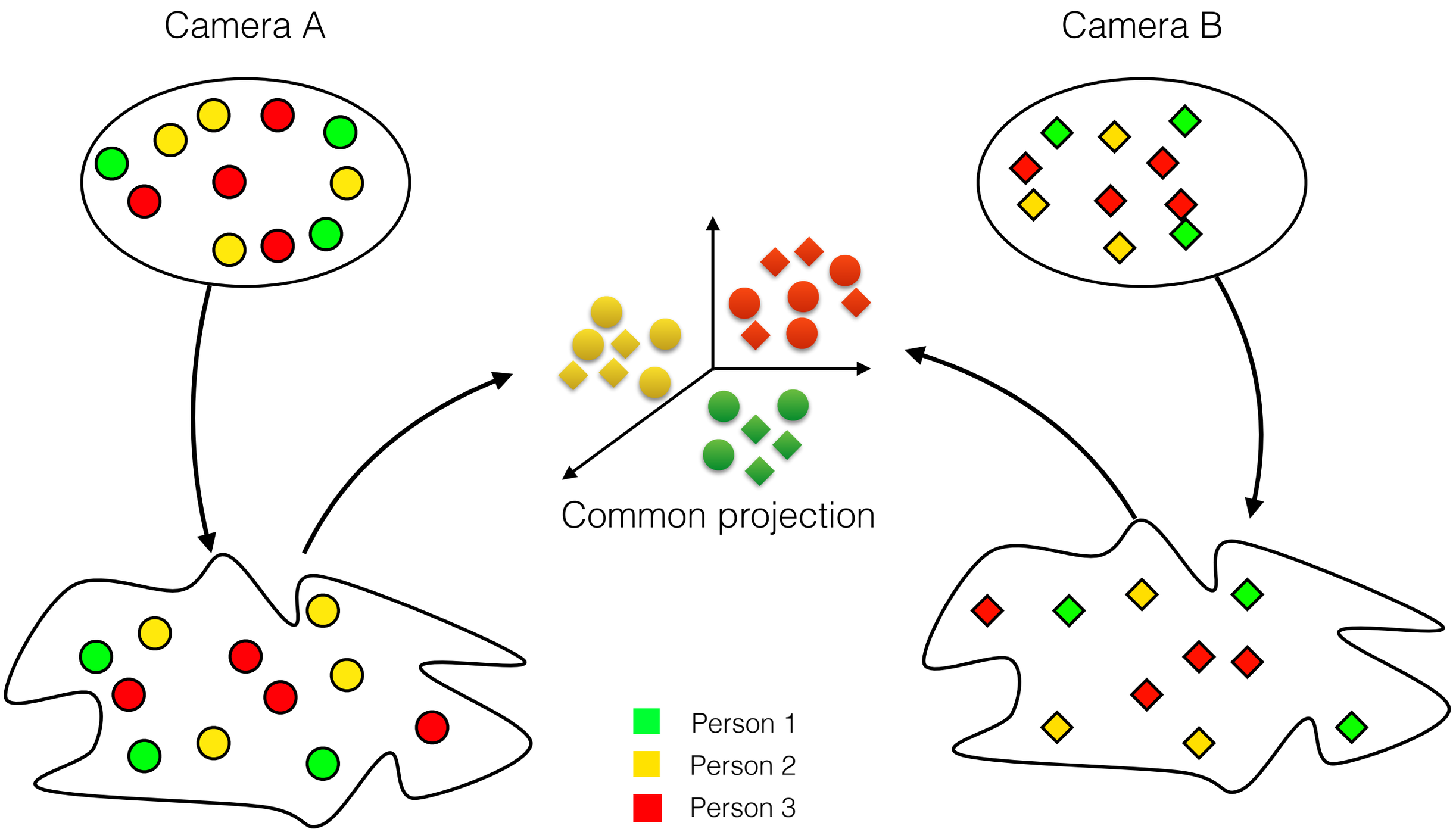
- Consider two cameras: compute the descriptor for all the person images, for each camera.

- Given a set of **paired** data from the two views the Canonical Correlation Analysis (CCA) find the projection directions that maximizes the correlation between the projected data:



Kernel Canonical Correlation Analysis

- the Kernel Canonical Correlation Analysis (KCCA) performs as CCA but on data projected through an opportune kernel (kernel trick to map descriptor)



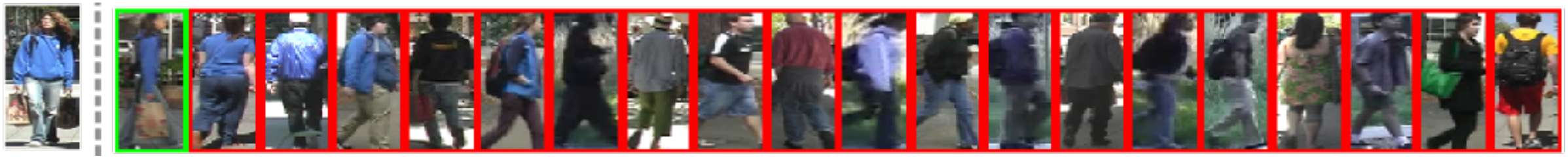
Re-Identification (NN)

- Project the gallery and probe samples using the weights learned with KCCA:

Probe

Gallery

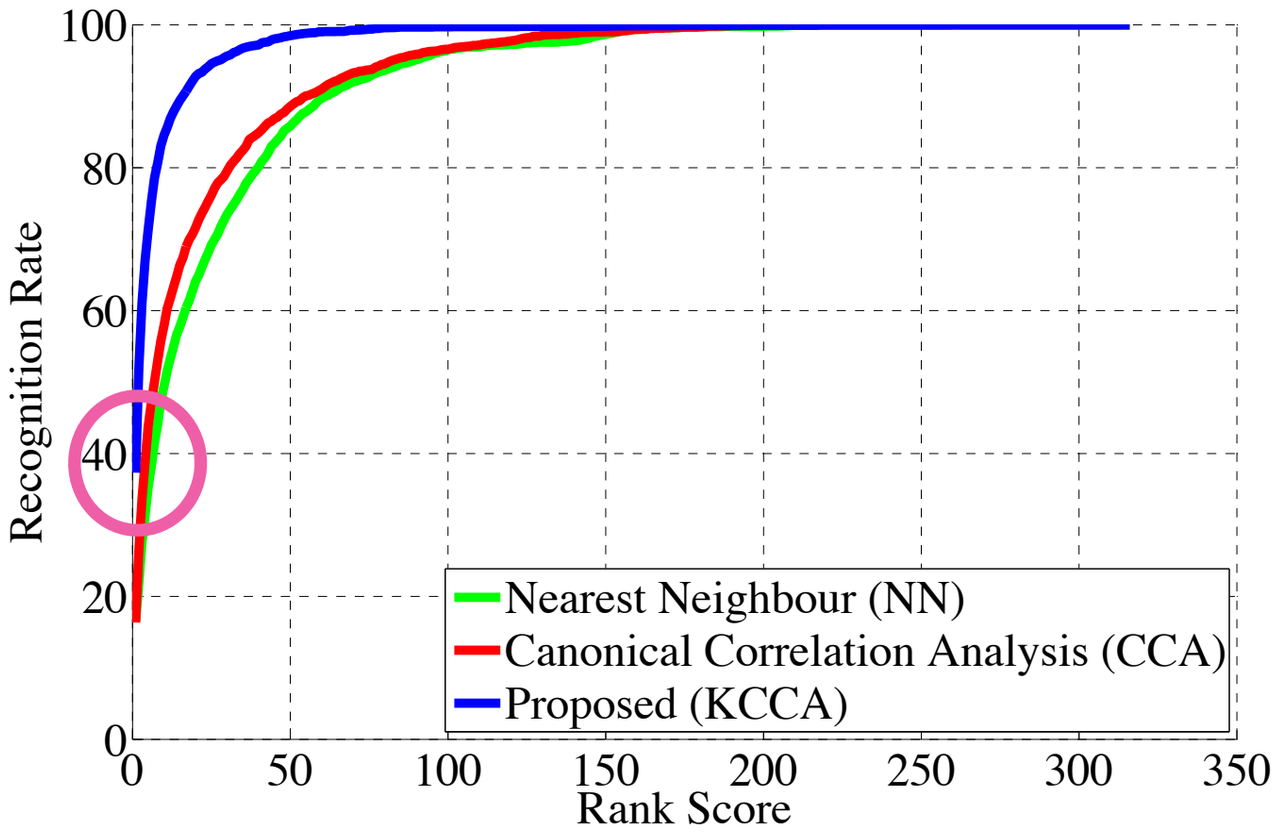
- Compute the cosine distance between the projected descriptors of the gallery and probe and perform a simple Nearest Neighbor (NN) classification:



Tests

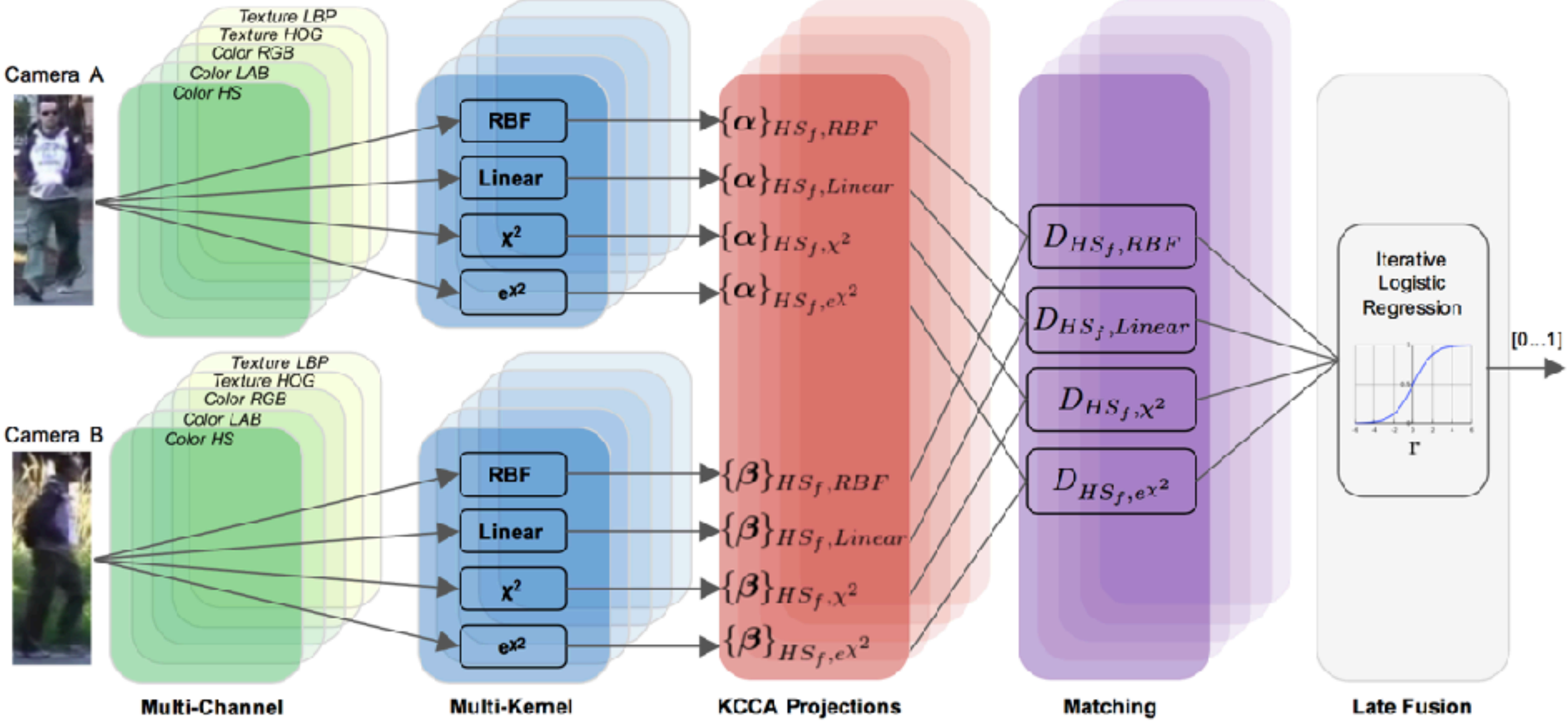
VIPER dataset

- 632 persons imaged by two non overlapping cameras
- Viewpoint changes up to 180 degrees, illumination changes and highlights.
- it is only appropriate for the single-shot re-identification modality.



	VIPeR				
Rank:	*1*	*10*	*20*	*50*	*100*
LMNN [20]	17	54	69	88	96
ITML [5]	13	53	71	90	97
PRDC [22]	16	54	70	87	97
DDC [10]	–	–	–	–	–
EIML [11]	22	63	78	93	98
ICT [2]	14	60	78	–	–
RPLM [12]	27	69	83	95	99
eSDC [19]	27	62	76	–	–
SalMatch [21]	30	65	–	–	–
Proposed	**37**	**85**	**93**	**98**	**100**

Multi Channel-Kernel Canonical Correlation Analysis for Cross-View Person Re-Identification



3D Morphable Model for Face Recognition

C. Ferrari

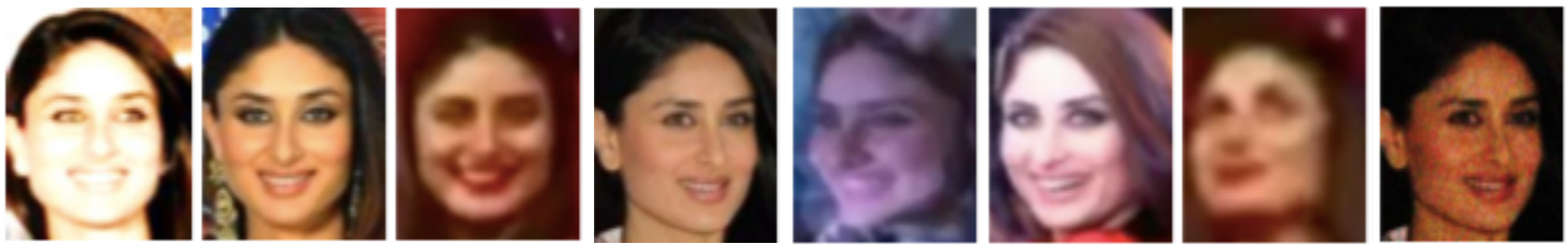
S. Berretti

A. Del Bimbo

Face Recognition

Given still images or videos, identify or verify a person's identity using a stored database of faces.

- The problem is still arduous when a face is captured "in the wild" i.e. under unconstrained conditions as for example strong variations in illumination, pose, expression, age etc.



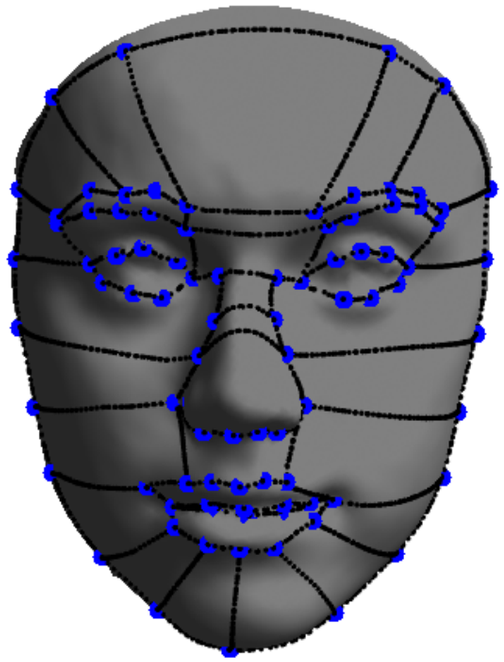
- 3D information can be exploited to lighten the negative effect of such issues and make the recognition module somewhat invariant.

- Applying **3D face modeling** techniques can further improve the performance of the 2D recognition task [Blanz, Vetter, '03]
  - Issue: 3D dense model registration

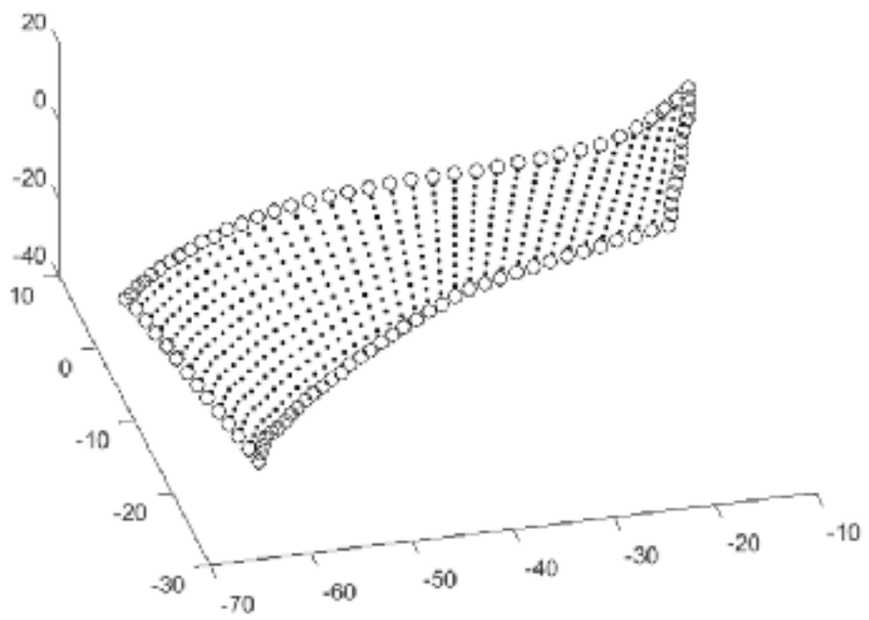
3D Dense Registration

This step is intended to reparametrize the 3D meshes so that corresponding points in all the scans have the same semantic meaning.

- **Face Partitioning**
  - Connect the 83 landmarks on each scan through geodesic paths to partition the face into 28 non overlapping closed regions.
  - Resample each geodesic path with a predefined geodesic distance.



- **Region Sampling**
  - Sample the interior surface of each region by connecting pairs of sampling points on opposite sides of a geodesic contour with a straight geodesic line.
  - In this way homologous regions are in dense correspondence across all the training scans.



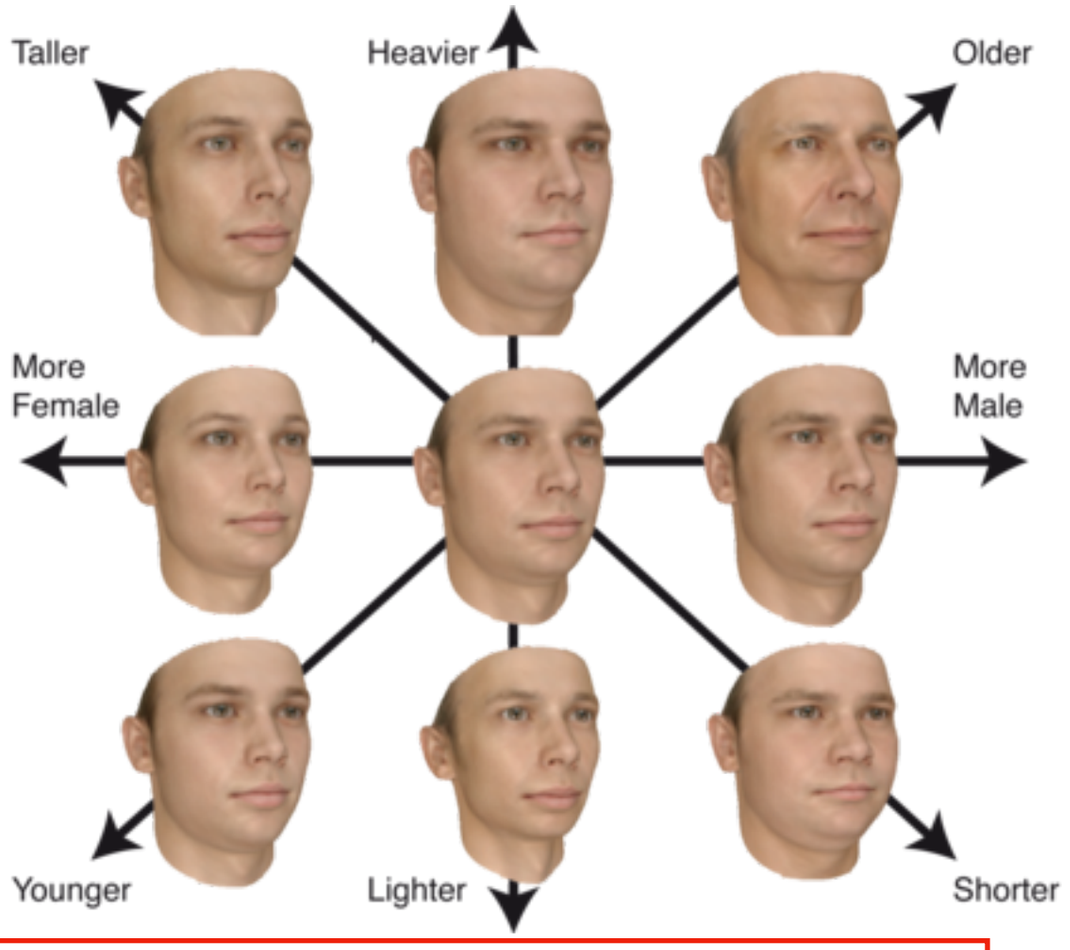
PCA based 3DMM

- The 3D Morphable Model (3DMM) is a technique that exploits the statistics of human faces to recover meaningful directions.

- Arbitrary new 3D shapes can be generated deforming an average 3D model through a linear combination of a set of direction vectors in the euclidean 3D space called deformation components.

- In the usual approach, they are recovered applying PCA on the aligned 3D models set. They represent the directions of maximum variability in the 3D face space.

- The great advantage of the 3DMM is that, even with a few control points, the whole model can be statistically morphed, since there exist movement relationships between sets of points.



Each component recovered with PCA affects the deformation of the whole shape; it is hard to model local variations without having to a priori segment the shape in regions and then interpolate them

Dictionary Learning based 3DMM

- Dictionary learning techniques aim at finding a dictionary **D** of k basis vectors whose linear combination best describes each of the n vectors in the training set.

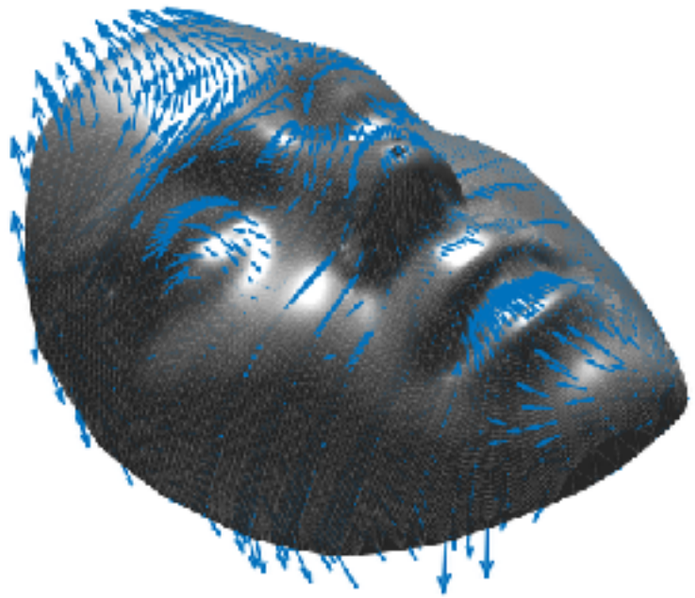
- We estimate the dictionary on the *vector field of the deviations* between each scan $\mathbf{f}_j$ and average model **m** so as to basically learn a dictionary of deformation directions.

The dictionary is estimated by minimizing the following, alternating between the two variables **D** (dictionary) and **w** (coefficients) [Mairal *et al,* '09]

$$\min_{\mathbf{w}_i \in \mathbb{R}^k,\, \mathbf{D} \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} \left(\|\mathbf{f}_i - \mathbf{D}\mathbf{w}_i\|_2^2 + \lambda_1 \|\mathbf{w}_i\|_1 + \lambda_2 \|\mathbf{w}_i\|_2 \right)$$

L1 sparsity inducing norm

L2 regularization



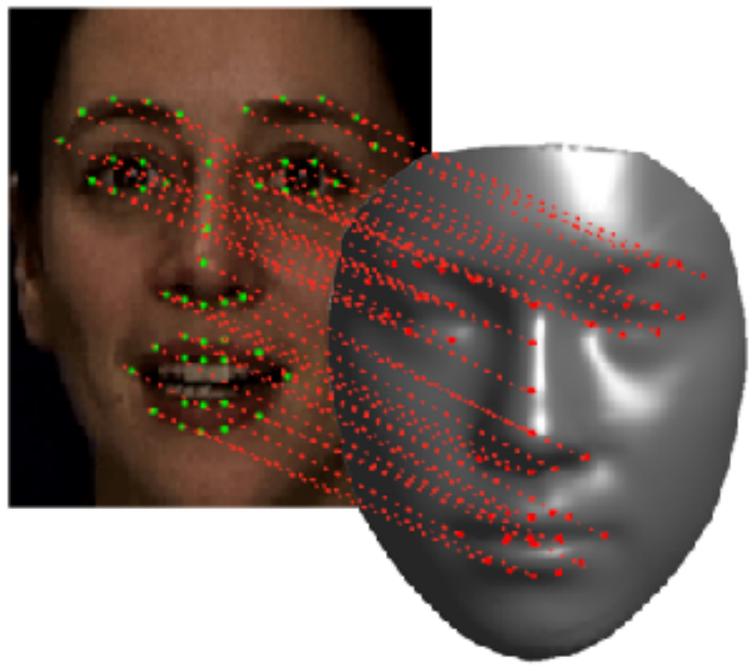
The average model and the deviation field of a real scan

PCA vs Dictionary Learning 3DMM

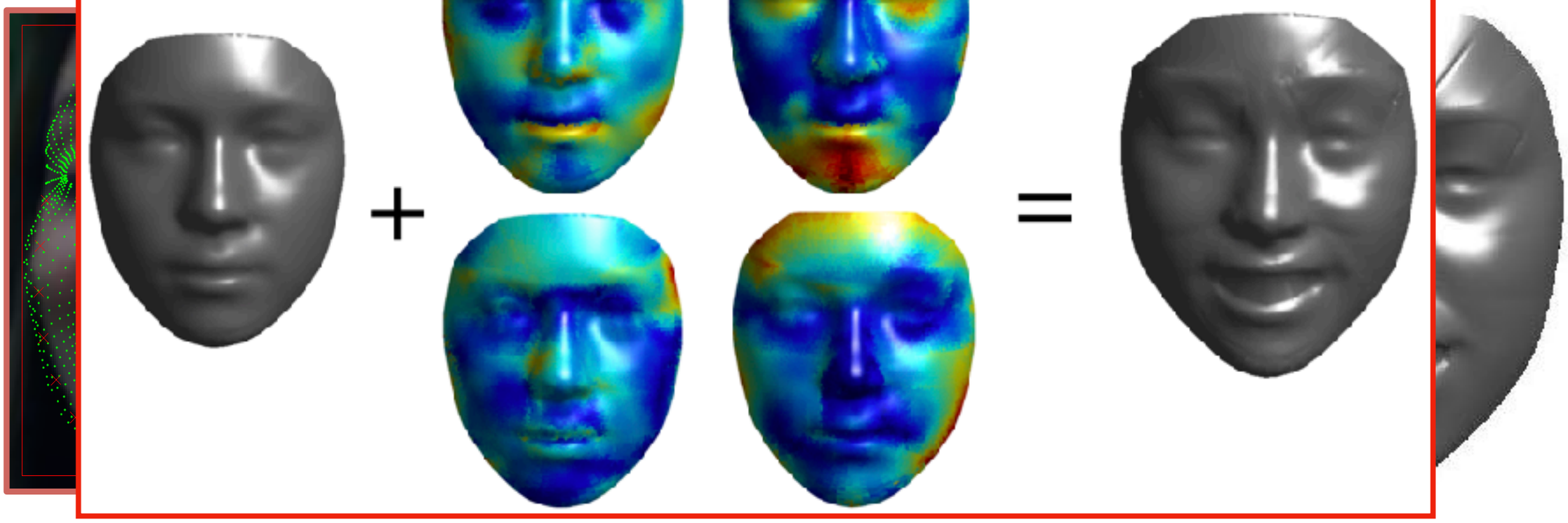
- The L1 norm used in the dictionary estimation leads to a shrinkage of many coefficients towards zero. This results in a sparsity effect and let us model:
  - the deformations locally
  - asymmetric deformations

Pose Estimation and Model Fitting

- Estimate the 3D head pose **P = [S, R, t]** by establishing a correspondence between a set of facial landmarks in 2D (automatically detected [De La Torre,'13]) and in 3D (manually labelled once) under an affine camera model



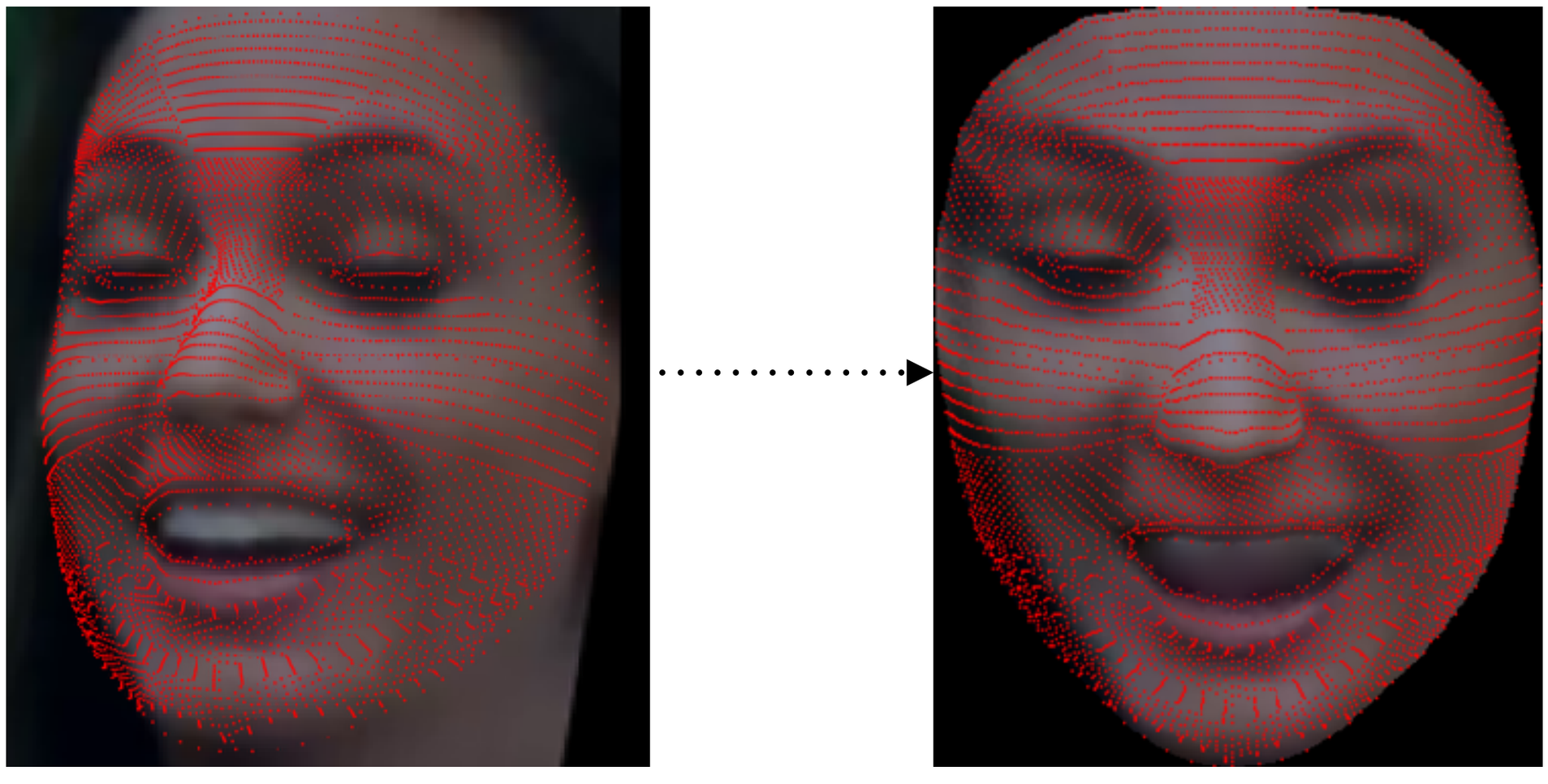
- The 3D
- reproj



Frontalization and Feature Extraction

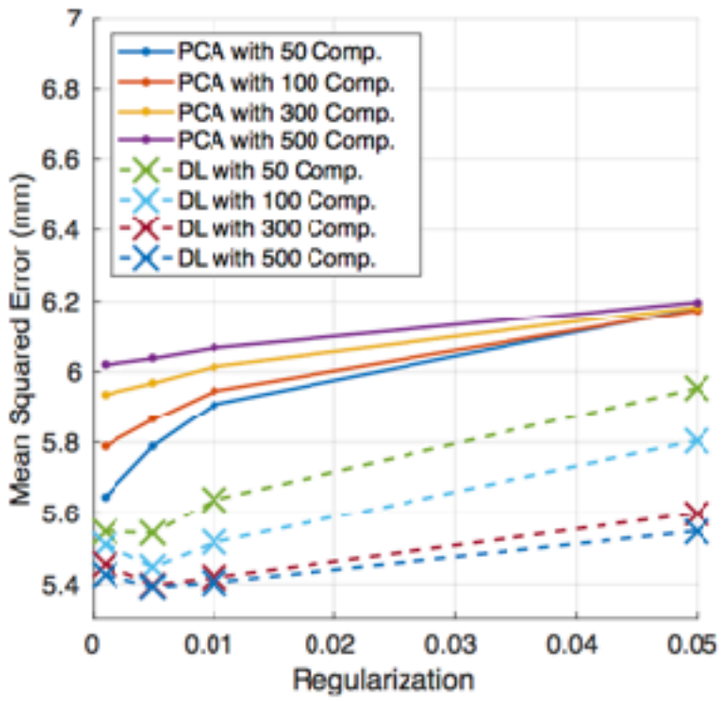
- Sample RGB values using the projected vertices locations on the original image to render a frontal view.

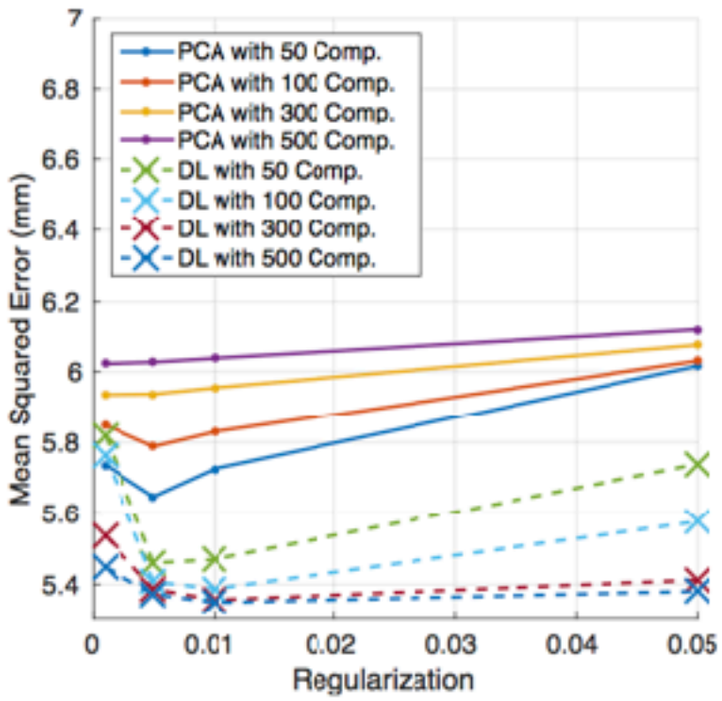
- Extract feature descriptors (LBP) on patches localized on the vertices projected on the rendered frontal view rather than on a regular grid

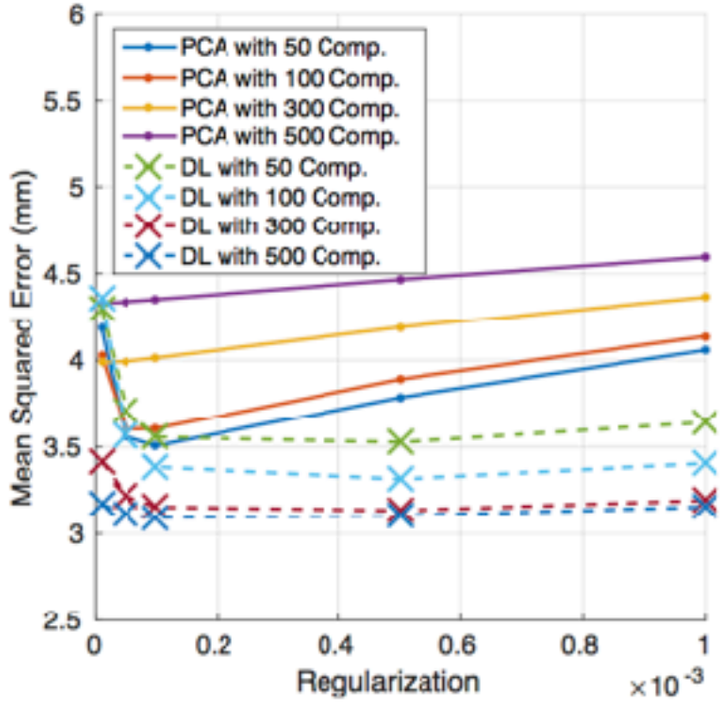


Tests

- 3D Reconstruction Error







- One single frontal neutral image per subject as gallery

- Probe images with strong variations in pose (+\- 45 yaw degrees) and expression

	#train 30		#train 50		#train 70	
	DL	PCA	DL	PCA	DL	PCA
Neutral	55.0	**55.9**	**56.1**	55.7	57.9	57.9
Angry	**46.3**	45.5	**47.6**	46.9	**49.5**	49.2
Disgust	**47.6**	44.9	**48.9**	46.7	49.7	49.7
Fear	**51.4**	49.7	**51.7**	49.9	**53.5**	53.2
Happiness	**51.0**	49.3	**50.5**	49.9	**53.8**	53.0
Sadness	**45.1**	44.8	**45.3**	44.6	48.5	**50.5**
Surprise	**39.3**	37.1	**40.4**	38.5	**41.7**	40.9
All	**47.2**	45.5	**47.9**	46.4	**49.8**	49.5

Investigating nuisance factors in face recognition with DCNN representation

Best Paper Award
(CVPRW on Biometrics)

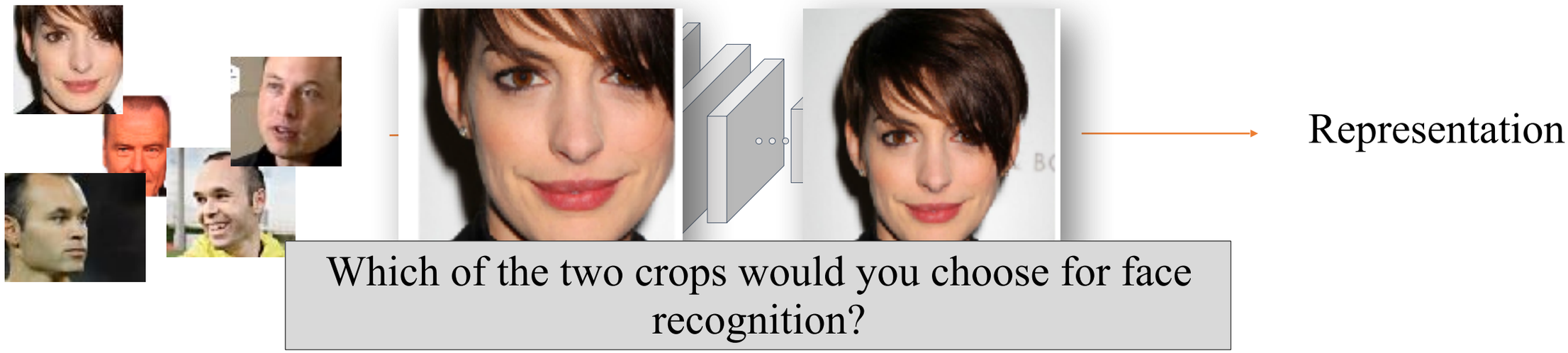
C. Ferrari

S. Berretti

A. Del Bimbo

Motivation

- DCNNs are very effective for many computer vision applications e.g. “face recognition in the wild ”;



- A powerful aspect of DCNNs is that they can learn image representation **from raw data end-to-end**, effectively.

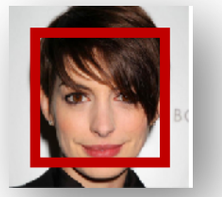
- On the other hand this fact can also be interpreted as a weakness, being **the network learning abilities totally dependent from the training data;**

- Preprocessing operations can tremendously affect the performance.

Goal

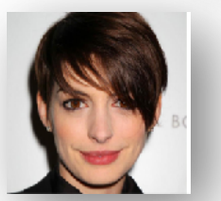
Evaluating the effect that different **data characteristics** and **pre-processing operations** have on the performance of DCNN architectures for face recognition;

- **Bounding box size:** trade-off between the face and background;

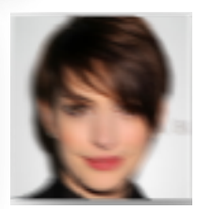


- **Alignment:** brings the faces into the same reference coordinate system

- **Data source:** whether the data is in the form of still images or video frames;

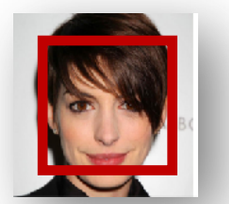


Still Image



Video Frame

- **Positioning:** is the relative position of the face inside the bounding box (when alignment is not applied).



Bounding Box Size

- The bounding box size defines the **visual content to be learned;**
- It can be also be interpreted as the **scale of the object** with respect to the context;
- This is crucial to let the network effectively identify the discriminative visual content for the specific task;

➢ **Tight**: it goes from the chin to just above the eyebrows. Reduces the background;

➢ **Large**: it includes at least the whole head. Increases the amount of useful information.

Alignment

The alignment is applied so that the same semantic content, e.g. the mouth, is in the same spatial position across all the images

- **2D Similarity:** the line connecting the eyes is horizontal. The distance between the eyes is 100px;

- **3D Frontalization:** exploits a 3D Morphable Model to compensate out-of-plane rotations of the head and render a frontal facing view.

Positioning

- Without alignment, the relative position of the face inside the bounding box can vary due to detection errors;

- Practically, **data augmentation** is applied to force invariance to translations and missing information;

➢ **Random Crop**: a random 224 x 224 crop is taken from the image

➢ **Upper half visible**: simulates the occlusion of the lower part of theface;

➢ **Lower half visible**: simulates the occlusion of the upper part of the face;

DCNN Architectures

The training data: VGGFace [2]. 2,622 different individuals with around 1,000 images each (2.6M)

Alexnet		
	Tight bounding box (**T**)	Large bounding box (**L**)
Aligned image (**AL**)	✓ (tr. from scratch)	✓ (tr. from scratch)
Original (not aligned) (**O**)	✓ (tr. from scratch)	✓ (tr. from scratch)
3D Frontalization (**F**)	✓ (tr. from scratch)	

VGGFace	
	Large bounding box (**L**)
Original (not aligned) (**O**)	(off-the-shelf)
3D Frontalization (**F**)	✓ (tr. from scratch)

[ParkhiBMVC2015]

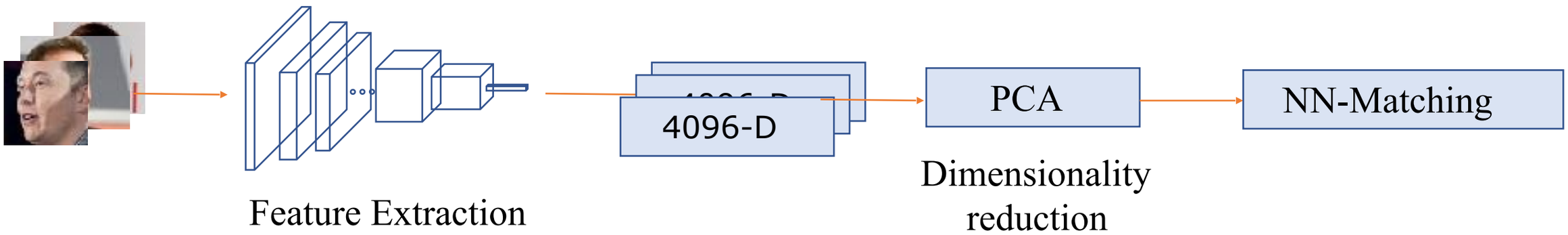
Experimental Results

- Experiments carried out on two benchmark datasets: **IJB-A** and **YouTubeFaces**;

  ➢ Both are divided in 10 splits for evaluation: 9 used as training set and 1 for testing;

  ➢ Face identification and verification accuracies are used as measures;

- A standard recognition pipeline was developed to assess the direct effect of image pre-processing operations;



Preprocessing Analysis

- Inconsistency between train and test data leads to a drop of performance;
- Best performance is obtained with non aligned images and larger bounding boxes.

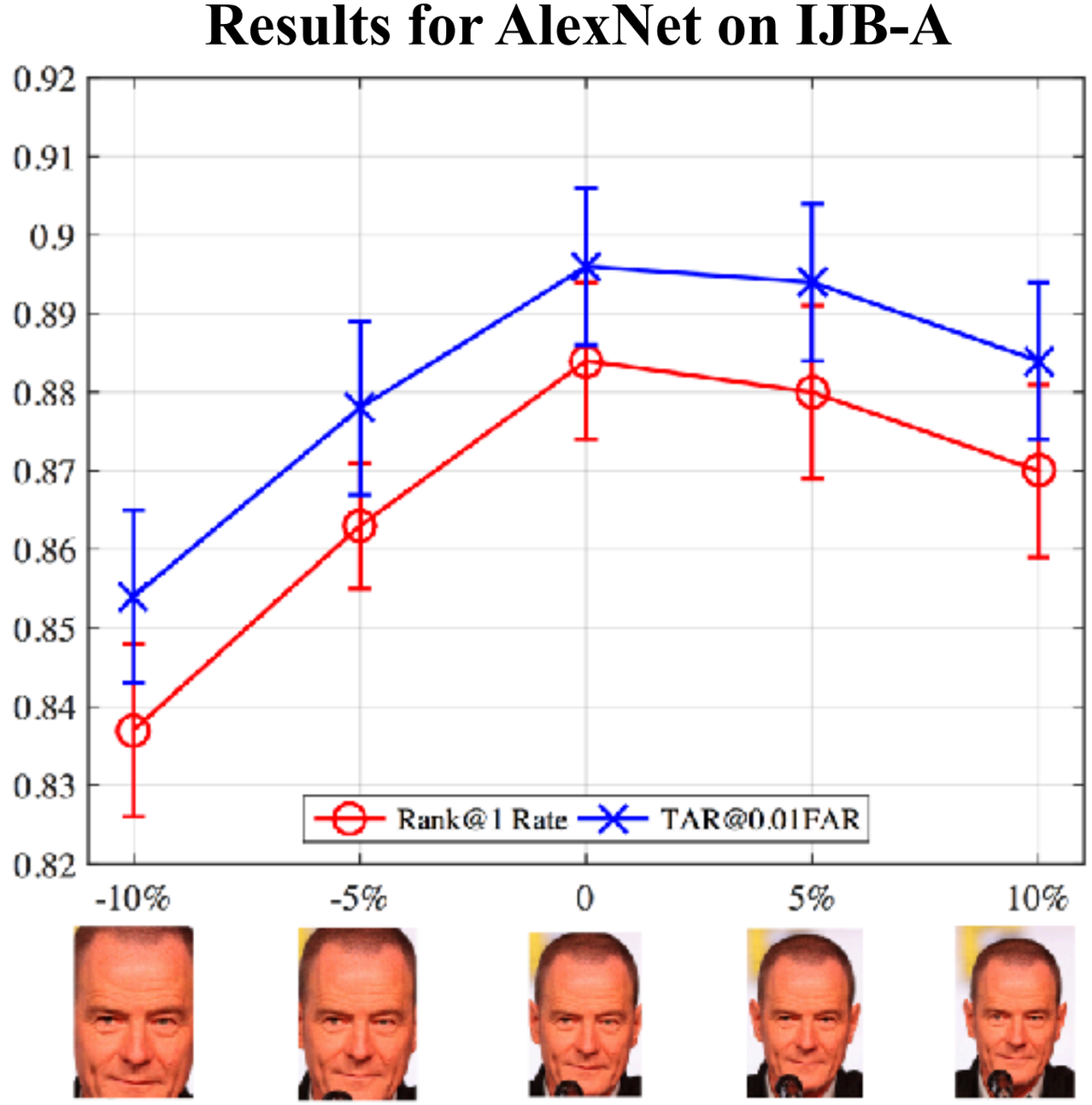
		Identification 1:N		Verification 1:1
Net	Data	TAR@0.01FAR	Rank@1	TAR@0.01FAR
AlexNet	AL-L	0.873 ± 0.012	0.861 ± 0.014	0.850 ± 0.018
AlexNet	O-L	$\mathbf{0.894 \pm 0.010}$	$\mathbf{0.886 \pm 0.010}$	$\mathbf{0.862 \pm 0.020}$
AlexNet	AL-T	0.827 ± 0.013	0.817 ± 0.016	0.808 ± 0.024
AlexNet	O-T	0.749 ± 0.020	0.750 ± 0.021	0.779 ± 0.024
AlexNet	F	0.839 ± 0.014	0.832 ± 0.019	0.817 ± 0.021

Table: Results on IJB-A for different data preprocessing methods

Preprocessing Analysis

Optimal Bounding Box Size:

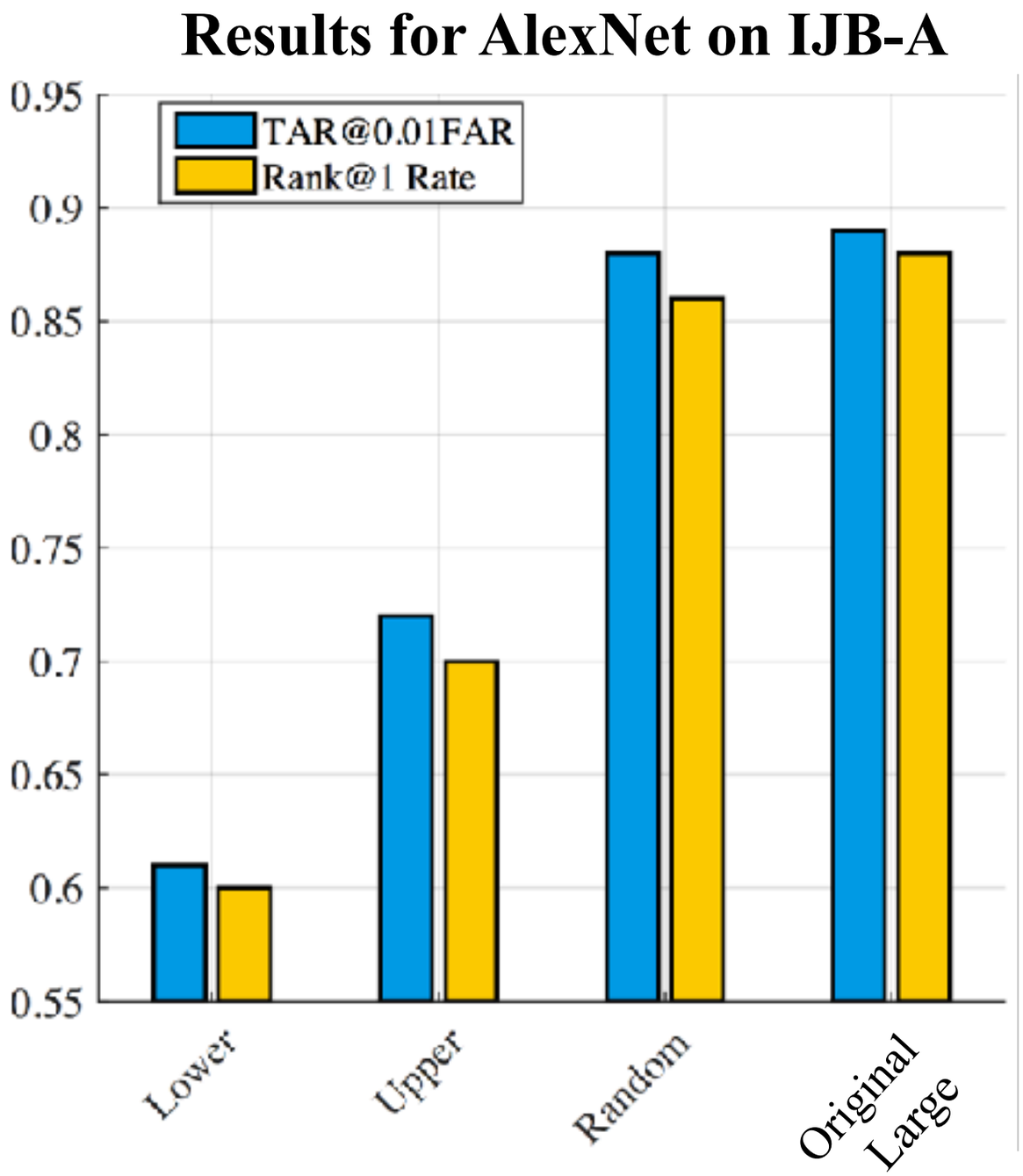
- ➢ Best results with a bounding box strictly including the whole head;
- ➢ Generally not straightforward to obtain such a bounding box;
- ➢ **Rule of thumb:** including background better than sacrificing content;



Preprocessing Analysis

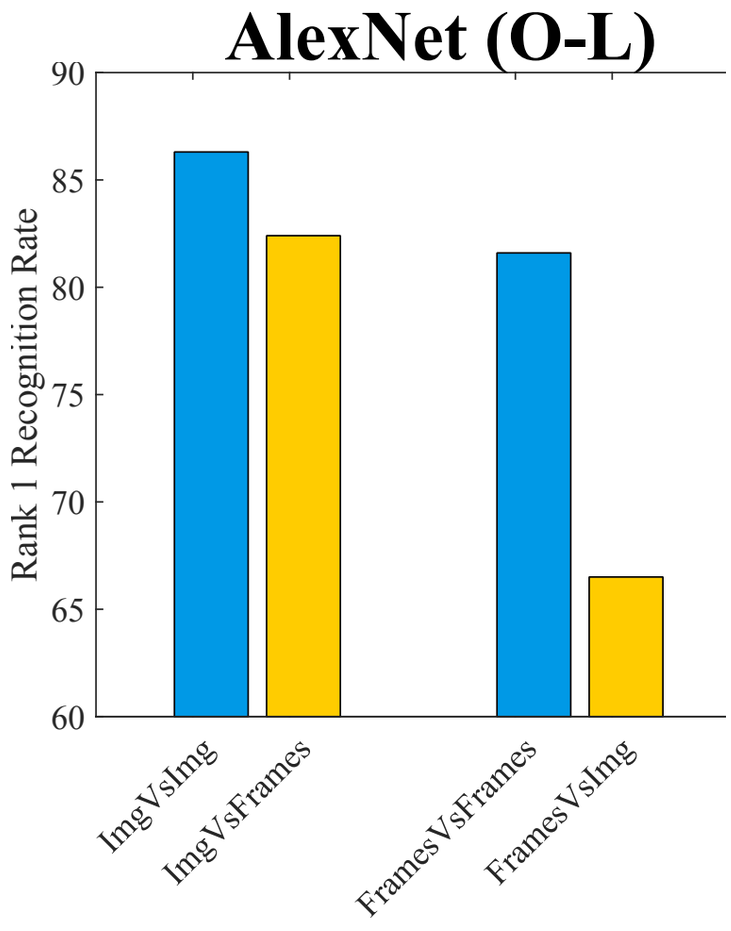
Positioning

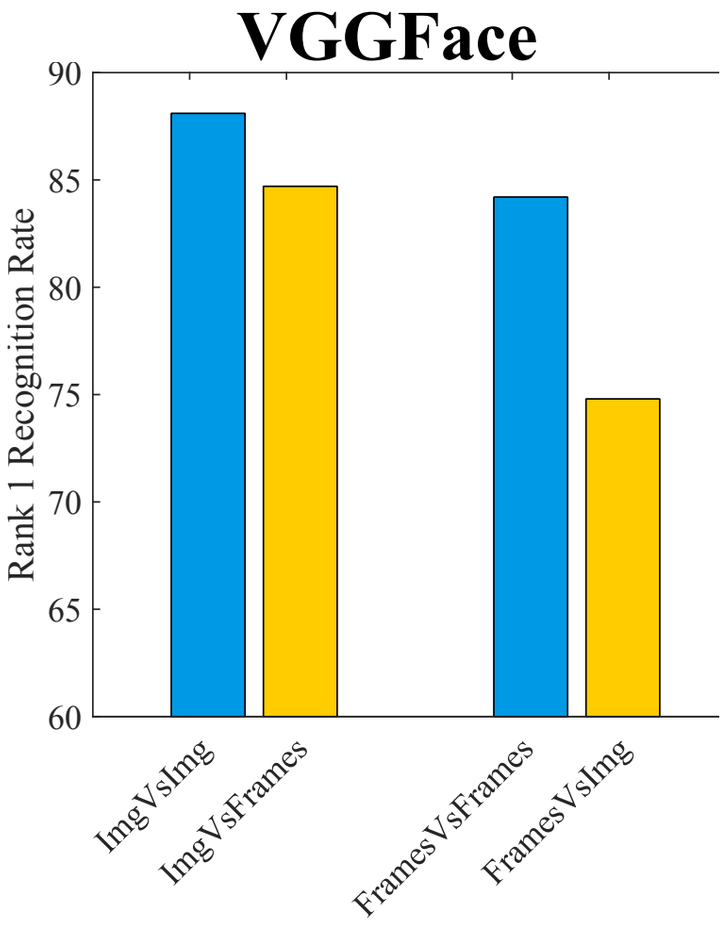
- ➢ Random crops don't affect performances because of data augmentation;
- ➢ The upper half of the face seems to carry more valuable information;



Data Source Analysis

- The IJB-A dataset includes both still images and video frames;
- A new protocol has been devised so as to perform matching between the two different data sources;
- Inconsistency between data sources negatively affects the performances.





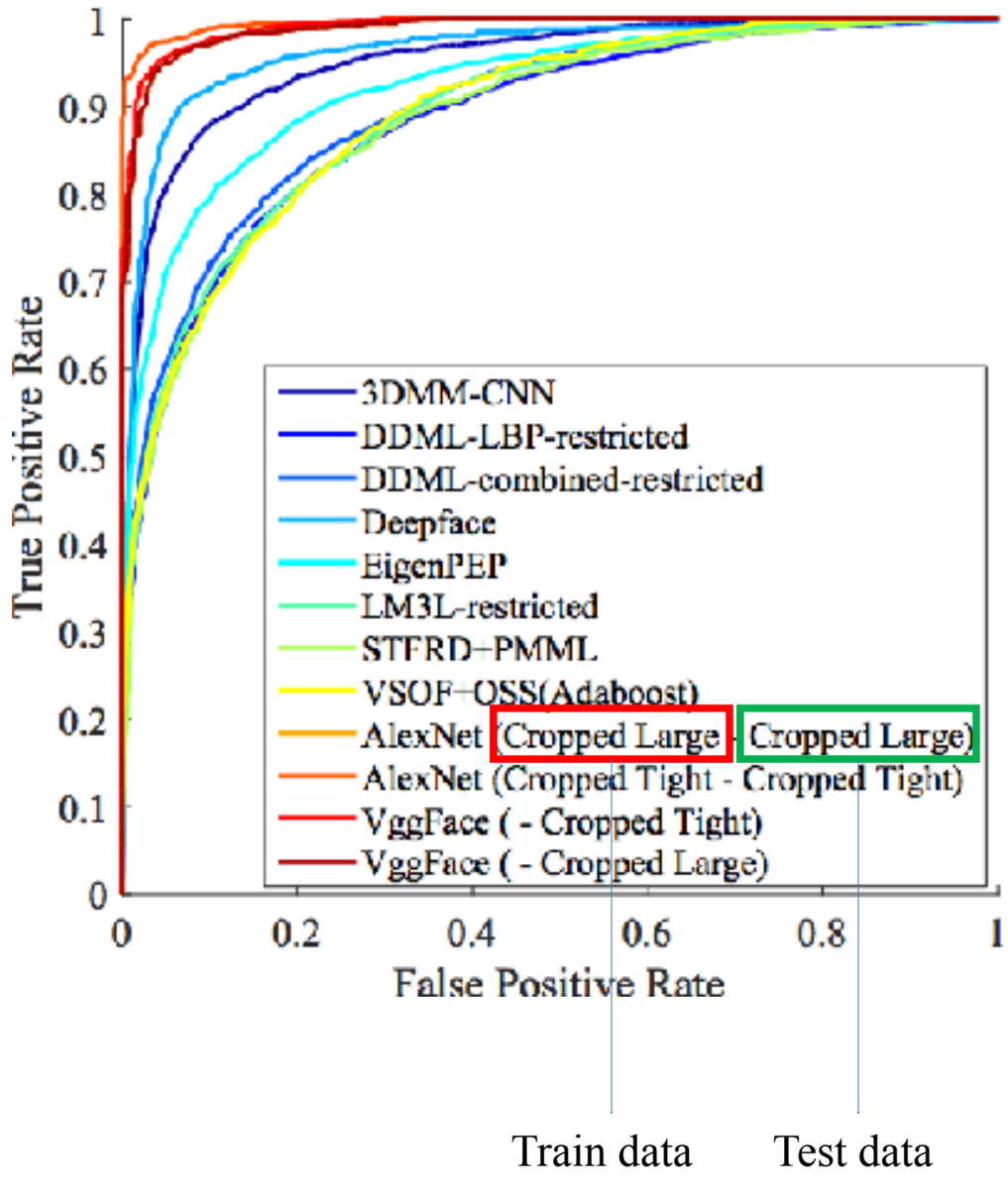
Still Image

Video Frame

[Gallery Vs Probe]

Comparison with State-of-the-Art: YTF

- **YouTubeFaces**: three network configurations have been considered: AlexNet (O-L), (O-T) and VggFace pre-trained;
- Test images are **not aligned** with both **large and tight crops**;
- Video based verification;
- Video descriptors computed as the average of the frames descriptors;
- **Best results** obtained with **AlexNet (O-T)**;
- Averaging the descriptors computed on images with some background could impair face recognition



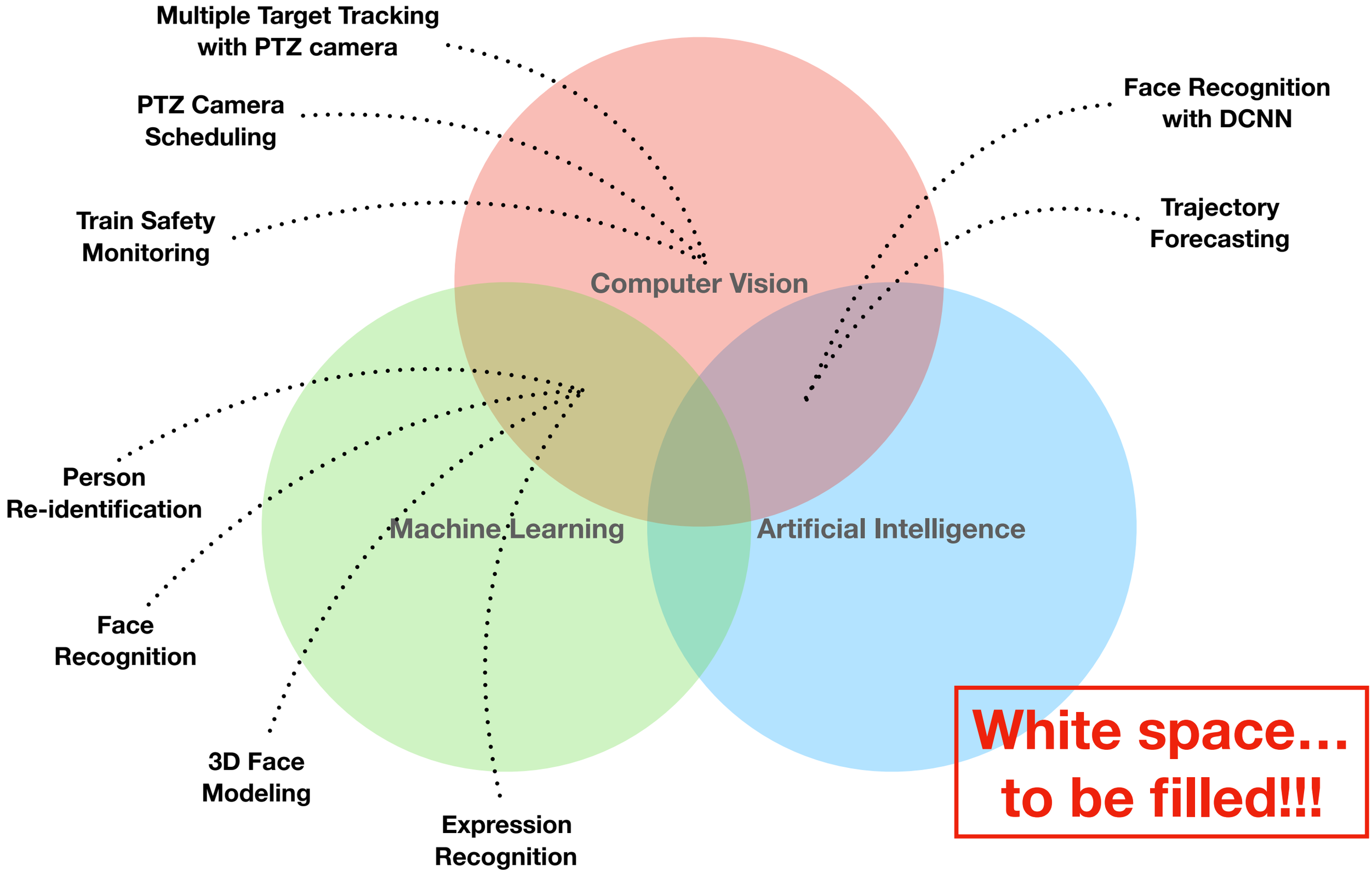
Discussion and conclusions

- **Consistency** between train and test data is fundamental.
- **Including background** is better than sacrificing content.
- **Learning from original data** is more effective than using aligned data.
- The **upper part** of the head i.e. eyes, eyebrows retains more discriminative identity information.
- **The source of the data plays an important role.** This could be ascribed to the different appearance distribution of the data.

Context-Aware Trajectory Prediction in Crowded Spaces

BMVC Submission #6

Research Interests



References

[LisantiTPAMI2015] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo, “Person re-identification by iterative re-weighted sparse ranking” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, iss. 8, pp. 1629-1642, 2015.

[LisantiTOMM2017] G. Lisanti, S. Karaman, and I. Masi, “Multi Channel-Kernel Canonical Correlation Analysis for Cross-View Person Re-Identification,” ACM Transactions on Multimedia, vol. 13, iss. 2, p. 13:1–13:19, 2017.

[KaramanPR2014] S. Karaman, G. Lisanti, A. D. Bagdanov, and A. Del Bimbo, “Leveraging local neighborhood topology for large scale person re-identification,” Pattern Recognition, vol. 47, iss. 12, pp. 3767-3778, 2014.

[LisantiICCV2017] G. Lisanti, N. Martinel, A. Del Bimbo, and G. Foresti, “Group Re-Identification via Unsupervised Transfer of Sparse Features Encoding,” in International Conference on Computer Vision, Venice, Italy, 2017.

[FerrariTMM2017] C. Ferrari, G. Lisanti, S. Berretti, and A. and Del Bimbo, “A Dictionary Learning based 3D Morphable Shape Model,” IEEE Transactions on Multimedia, 2017.

[BlanzVetterTPAMI2003] V. Blanz, and T. Vetter. "Face recognition based on fitting a 3D morphable model." IEEE Transactions on pattern analysis and machine intelligence 25.9 (2003): 1063-1074.

[FerrariICPR2016] C. Ferrari, G. Lisanti, S. Berretti, and A. Del Bimbo, “Effective 3D Based Frontalization for Unconstrained Face Recognition,” in Proceedings of the International Conference on Pattern Recognition, Cancun, Mexico, 2016.

[FerrariCVPRW2017] C. Ferrari, G. Lisanti, S. Berretti, and A. Del Bimbo, “Investigating Nuisance Factors in Face Recognition with DCNN Representation,” in Proceedings of the International Conference on Computer Vision and Pattern Recognition (Workshop on Biometrics, Honolulu, Hawaii, USA, 2017.

[ParkhiBMVC2015] Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. "Deep Face Recognition." BMVC. Vol. 1. No. 3. 2015.