Exploiting feature representations through similarity learning and ranking aggregation for person re-identification

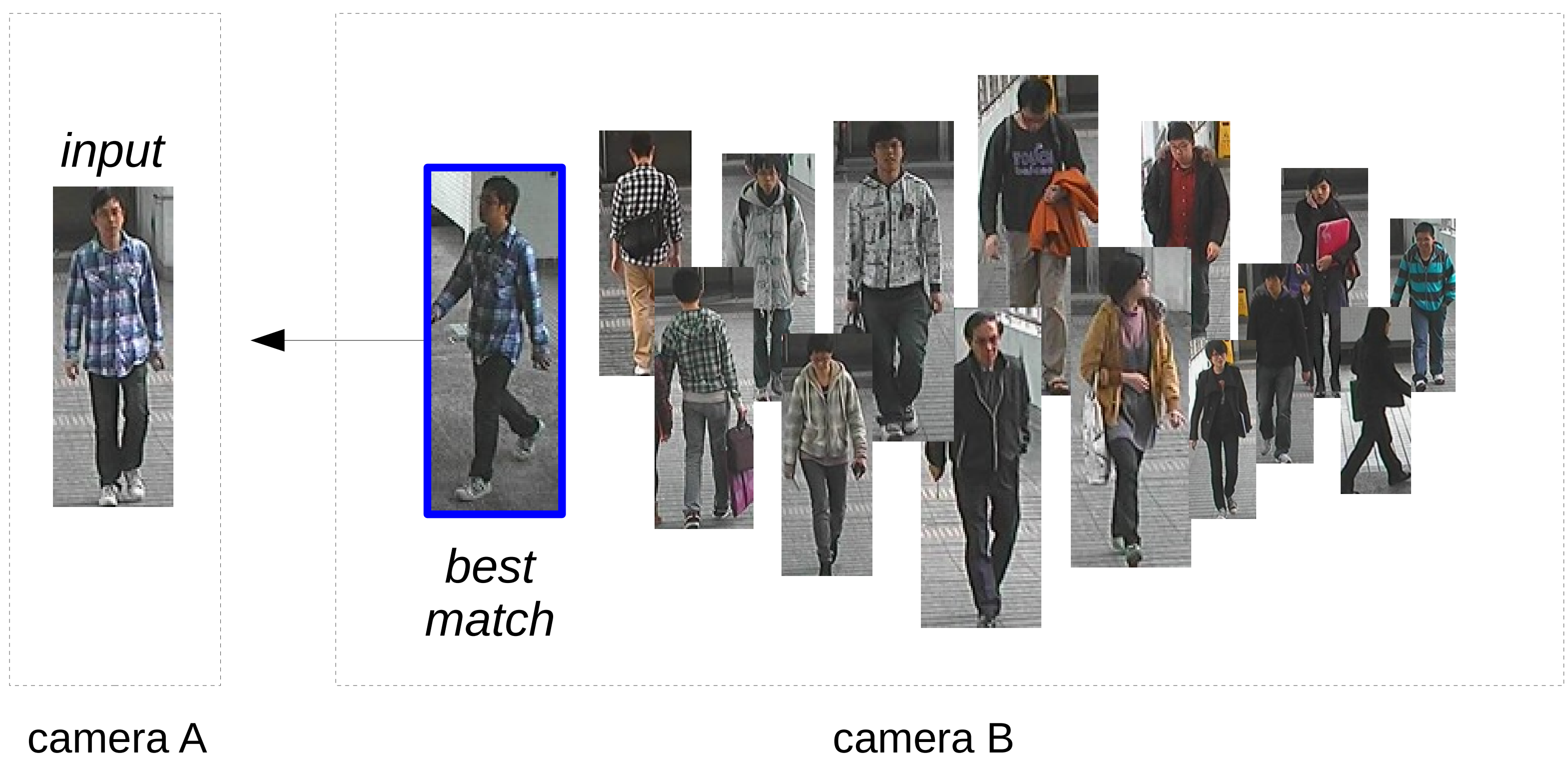
Julio C. S. Jacques Junior [1,2]
Xavier Baró [2,3]
Sergio Escalera [1,2]

[1] *Department of Mathematics and Informatics* - University of Barcelona, Spain
[2] *Computer Vision Center* - Universitat Autònoma de Barcelona, Spain
[3] *Faculty of Computer Science, Multimedia and Telecommunication* - Universitat Oberta de Catalunya, Spain

FG2017 - 12th IEEE Conference on Automatic Face and Gesture Recognition

Introduction

- The goal of **person re-identification** models is to retrieve the correct match, given a source image of a particular individual, from a large database.
  - Captured from different cameras, views and time intervals.



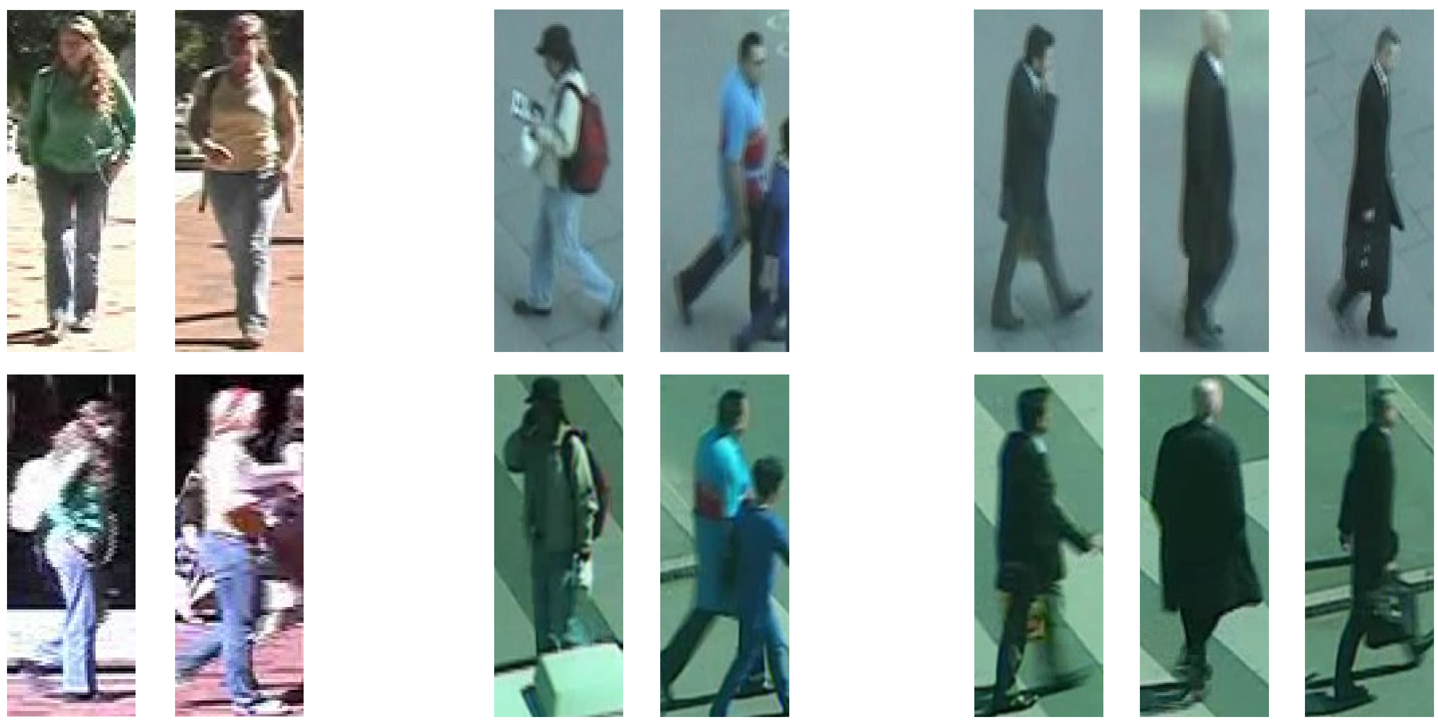
Introduction

- This task still presents main **open challenges**.

Illumination, pose variations | *camera settings, occlusions* | *strong visual similarity, etc*

camera A

camera B



Different strategies

- **Feature representation:** construct robust and discriminative features in order to describe the appearance of the same individual across different camera views [5],[9],[10],[11].

- **Distance metric learning:** to learn a metric in the image feature space that keep features coming from same class closer, while features from different classes farther apart [7],[12].

Different strategies

- ***Domain adaptation:*** to address the view-specific feature distortion problem (*transfer learning*) [24].

- ***Convolutional Neural Networks (CNN):*** provide a powerful and adaptive tool without excessive usage of hand-crafted features [4],[9],[11],[14],[25].

Concatenation of hand-crafted features *sometimes would be more distinctive and reliable (Wu et al. [9]).*

[9] S. Wu, Y. C. Chen, X. Li, A. C. Wu, J. J. You, and W. S. Zheng, "An enhanced deep feature representation for person re-identification," in IEEE Winter Conf. on Applications of Computer Vision (WACV), 2016.

Proposed Model

- Exploit **different feature representations** through the combination of ***new and complementary features*** within the framework proposed by Chen et al.[5], followed by a ranking aggregation strategy.
  - Enforces **similarity learning metric** (built on the recently proposed *polynomial feature map* [7])
  - With **spatial constraints**.

[5] D. Chen, Z. Yuan, B. Chen, and N. Zheng, “Similarity learning with spatial constraints for person re-identification,” in CVPR, 2016.

[7] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, “Similarity learning on an explicit polynomial kernel feature map for person re-identification,” in CVPR, 2015.

Proposed Model

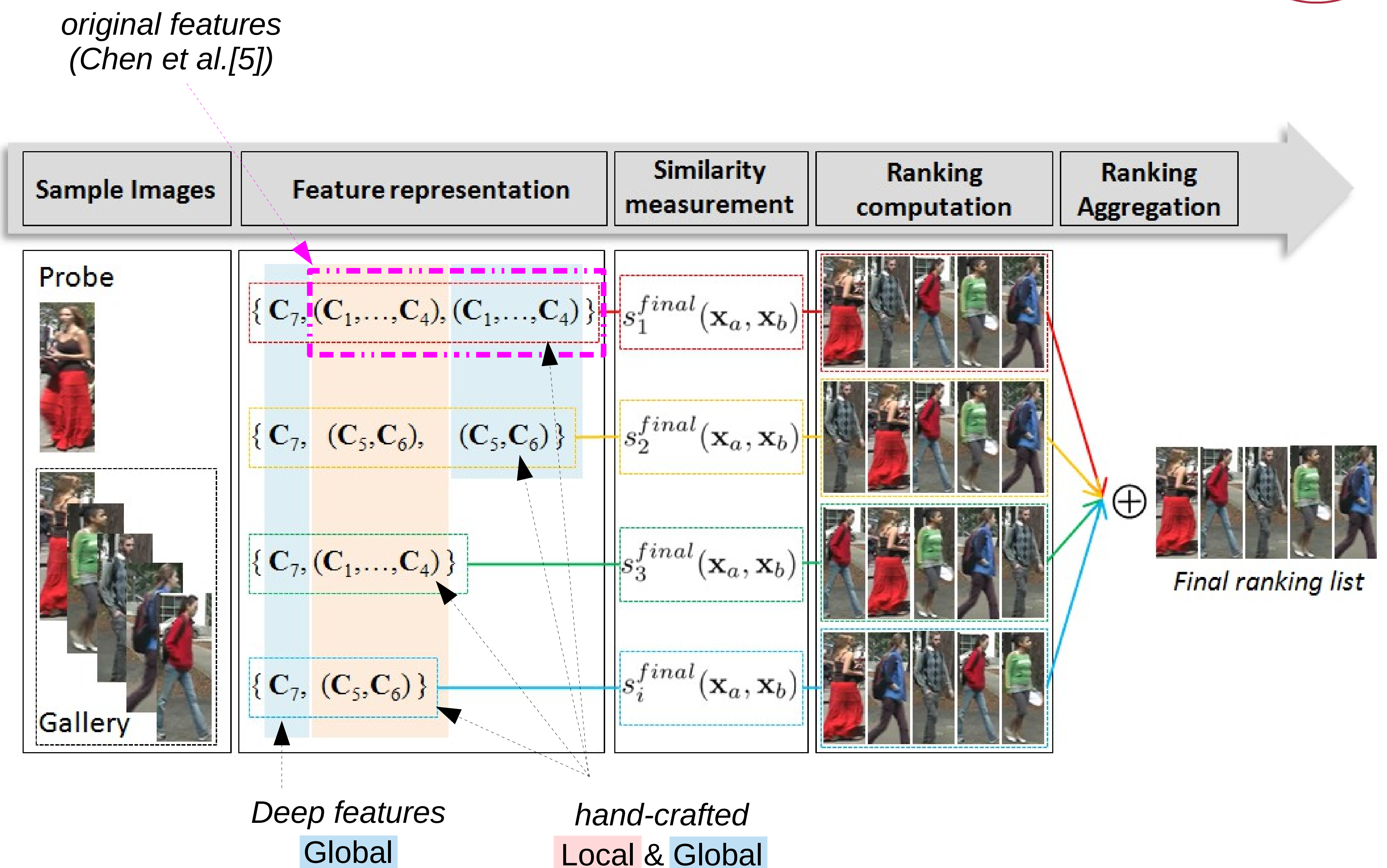
- Exploit **different feature representations** through the combination of ***new and complementary features*** within the framework proposed by Chen et al.[5], followed by a ranking aggregation strategy.
  - Enforces **similarity learning metric** (built on the recently proposed *polynomial feature map* [7])
  - With **spatial constraints**.

- We **advanced the state-of-the-art** on *VIPeR* and *PRID450s* datasets (by 8.89% and 6.9%, respectively) and obtained competitive results on CUHK01 database.

[5] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in CVPR, 2016.

[7] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in CVPR, 2015.

Proposed Model



Polynomial Feature Map

- In order to *measure the similarity between image descriptors* $\mathbf{x}_a, \mathbf{x}_b \in \mathbb{R}^{d\times 1}$, we learn the **similarity function** as:

$$f(\mathbf{x}_a, \mathbf{x}_b) = \langle \phi(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W} \rangle_F,$$

where $\langle \cdot, \cdot \rangle_F$ is the *Frobenius* inner product.

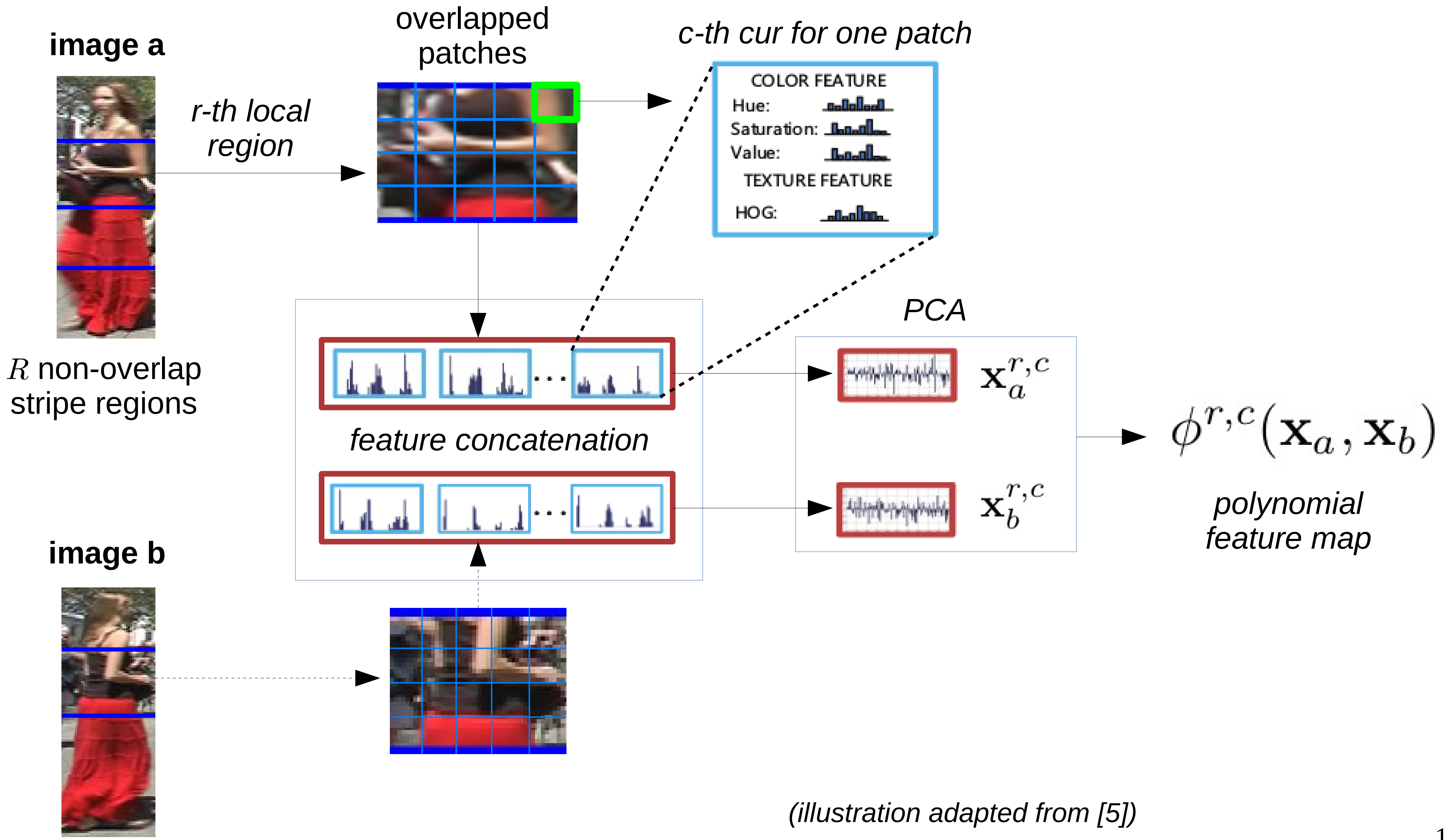
$$f(\mathbf{x}_a, \mathbf{x}_b) = \underbrace{\langle \phi_M(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}_M \rangle_F}_{(\mathbf{x}_a - \mathbf{x}_b)^\top \mathbf{W}_M (\mathbf{x}_a - \mathbf{x}_b)} + \underbrace{\langle \phi_B(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}_B \rangle_F}_{\mathbf{x}_a^\top \mathbf{W}_B \mathbf{x}_b + \mathbf{x}_b^\top \mathbf{W}_B \mathbf{x}_a}$$

Mahalanobis distance

Bilinear similarity

Feature map dimension is reduced by means of PCA for $\mathbf{x}_a$ *and* $\mathbf{x}_b$ *before its generation.*

Spatially Constrained Similarity Function



(illustration adapted from [5])

Local similarity integration

- In order to **combine multiple visual cues** *within a local region*, the following linear similarity function is employed :

$$s^r(\mathbf{x}_a, \mathbf{x}_b) = \sum_{c=1}^{C} \langle \phi^{r,c}(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}^{r,c} \rangle_F,$$

where $\mathbf{W}^{r,c} = [\mathbf{W}_M^{r,c}, \mathbf{W}_B^{r,c}]$ and $\mathbf{W}_M^{r,c}$, $\mathbf{W}_B^{r,c}$ correspond to $\phi_M^{r,c}(\mathbf{x}_a, \mathbf{x}_b)$ and $\phi_B^{r,c}(\mathbf{x}_a, \mathbf{x}_b)$, respectively.

- *Local similarity* are **integrated** as follows:

$$s^{local}(\mathbf{x}_a, \mathbf{x}_b) = \sum_{r=1}^{R} s^r(\mathbf{x}_a, \mathbf{x}_b)$$

Global-Local collaboration

- To describe the matching of **large patterns**, the polynomial feature map is also used for the whole image

$$s^{global}(\mathbf{x}_a, \mathbf{x}_b) = \sum_{c=1}^{C} \langle \phi^{G,c}(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}^{G,c} \rangle_F,$$

where $\mathbf{W}^{G,c} = [\mathbf{W}_M^{G,c}, \mathbf{W}_B^{G,c}]$ and $\mathbf{W}_M^{G,c}$, $\mathbf{W}_B^{G,c}$ correspond to $\phi_M^{G,c}(\mathbf{x}_a, \mathbf{x}_b)$ and $\phi_B^{G,c}(\mathbf{x}_a, \mathbf{x}_b)$, respectively.

- Finally, **local and global** similarity functions are **combined** and the *<u>overall similarity score</u>* is given by:

$$s(\mathbf{x}_a, \mathbf{x}_b) = s^{local}(\mathbf{x}_a, \mathbf{x}_b) + \gamma s^{global}(\mathbf{x}_a, \mathbf{x}_b),$$

where $\gamma = 1.1$

Visual Cues

- Chen et al.[5] proposed to use four visual cues, *extracted for each patch/region*:

  - ***Joint*** [1] (8x8x8) *and* ***concatenated*** [2] (48 bin) *histograms:* HSV and LAB
  - HOG and SILTP (*Scale Invariant Local Ternary Pattern*)

  - C_1 = HSV_1/HOG
  - C_2 = HSV_2/SILTP
  - C_3 = LAB_1/SILTP
  - C_4 = LAB_2/HOG

- PCA is applied to reduce dimensionality (*d=120*)
- Normalized

Complementary features

- We propose to include *<u>new and complementary features</u>* within the similarity function presented in [5]
  - **SCNCD [6]:** indicates the probability of a color being assigned to several nearest color names.
    - Extracted from RGB, *normalized rgb*, $l_1l_2l_3$, and HSV, and fused.
  - **Context information:** image-foreground feature representation [6], based on a Deep Decompositional Network (DDN) [21].

(Locally & Globally)

Color & **Texture** integration

C_5 = SCNCD/HOG

C_6 = SCNCD/SILTP

PCA + Normalization

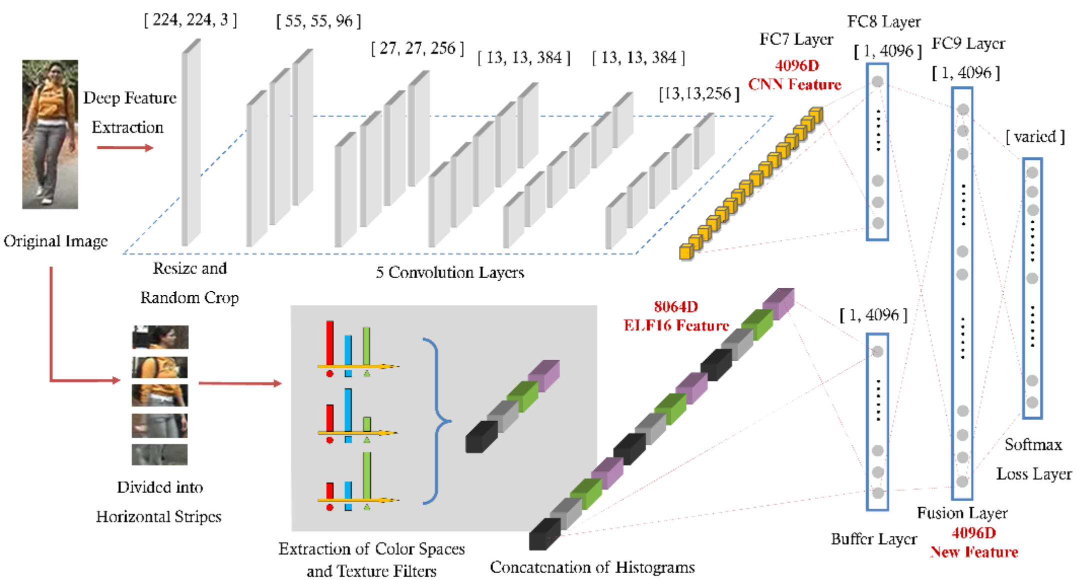
[6] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, “Salient color names for person re-identification,” in ECCV, 2014.

[21] P. Luo, X. Wang, and X. Tang, “Pedestrian parsing via deep decom-positional network,” in ICCV, 2013.

Complementary features

– **Deep Features:** *Feature Fusion Net* (Wu et al. [9]).

- *CNN and hand-crafted* features are combined to **produce an image description from the last convolutional layer.**
- Finally, *Deep feat* + *LOMO* *(Local Maximal Ocurrence)* demonstrated to have **higher discriminative power** (31056D feature vector).



(image from [9])

+ LOMO

C_7 = *Deep Feature*

PCA + Normalization

[9] S. Wu, Y. C. Chen, X. Li, A. C. Wu, J. J. You, and W. S. Zheng, “An enhanced deep feature representation for person re-identification,” in IEEE Winter Conf. on Applications of Computer Vision (WACV), 2016.

Integration strategy

- We compute **4 similarity measures** using different descriptors, in order to obtain **complementary ranking lists**.

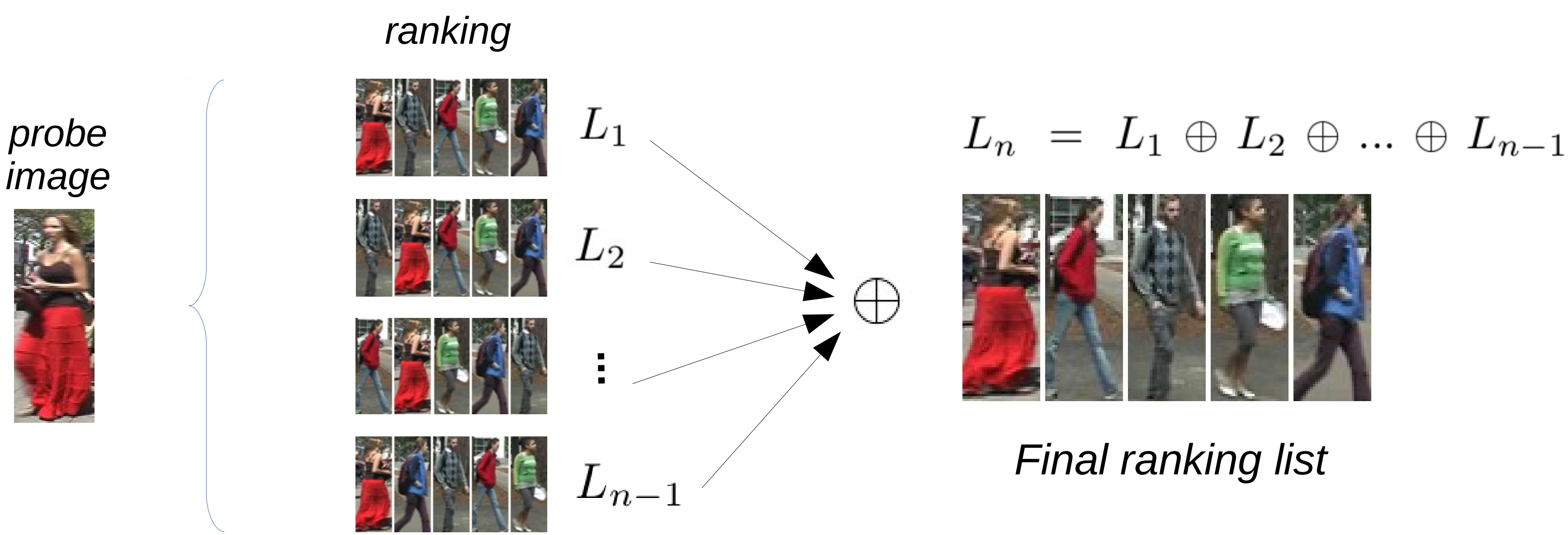
Features	Type	Local	Global
baseline	F_0	C_1 to C_4	C_1 to C_4
baseline + deep feat.	F_1	C_1 to C_4	C_1 to C_4, C_7
SCNCD + context + deep feat.	F_2	C_5, C_6,	C_5, C_6, C_7
simplified version of F_1	F_3	C_1 to C_4	C_7
simplified version of F_2	F_4	C_5, C_6,	C_7

$s_i^{final}(\mathbf{x}_a, \mathbf{x}_b)$,

$i \in \{1, 2, 3, 4\}$

Ranking aggregation

- Different ranking lists are generated and combined, using the Stuart ranking aggregation [23].
  - It is a probabilistic method based on order statistics.
  - Our goal is to improve accuracy by exploiting different feature representations that may complement each other.



[23] *J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," Science, 2003.*

Experimental results

- **Three case studies**, on three broadly employed public datasets: VIPeR, PRID450s and CUHK01.
  - State-of-the-art comparison
  - Influence of context information within SCNCD
  - Accuracy performance obtained by each complementary feat.

- Using a well known evaluation protocol (**single shot scenario**)
  - 50% training and 50% testing, without overlap on person identities
  - Camera A → probe set and Camera B → gallery set.
  - Each probe image is matched against every gallery set image and the rank of correct match obtained
  - Average of *Cumulative Matching Characteristic* (CMC) curves across 10 partition is reported.

Case 1:
State-of-the-art comparison

Rank	1	5	10	20
VIPeR				
Our	**58.77**	**86.39**	**93.48**	**97.82**
SCSP [5]	53.54	82.59	91.49	96.65
Deep+LOMO [9]	51.06	81.01	91.39	96.90
TCP [4]	47.80	74.70	84.80	91.10
CMC [13]	45.90	77.50	88.90	95.80
Mirror [10]	42.97	75.82	87.28	94.84
LSSCDL [24]	42.66	-	84.27	91.93
FT-JSTL+DGD[11]	38.60	-	-	-
CBRA [16][7]	31.20	60.80	74.30	85.90

baseline (SCSP [5])
FFN (deep feat.) (Deep+LOMO [9])

PRID450s				
Our	**71.56**	**90.58**	**94.40**	**96.98**
Deep+LOMO [9]	66.62	86.84	92.84	96.89
LSSCDL [24]	60.49	-	88.58	93.60
Mirror [10]	55.42	79.29	87.82	93.87
CBRA [16][7]	26.40	57.10	71.00	83.20

FFN (deep feat.) (Deep+LOMO [9])

New features demonstrated to **complement each other**, being very powerful.

Case 1: State-of-the-art comparison

Rank	1	5	10	20
CUHK01				
FT-JSTL+DGD[11]	**66.60**	-	-	-
LSSCDL [24]	65.97	$\approx$ **88.0**	$\approx$ **92.0**	$\approx$ 96.0
Our	59.63	83.66	89.71	94.39
Deep+LOMO [9]	55.51	78.40	83.68	92.59
3TCP [4][8]	53.70	84.30	91.00	**96.30**
CMC [13]	53.40	76.40	84.40	90.50
Mirror [10]	40.40	64.63	75.34	84.08

- Xiao et al. [11] was designed to learn features from multiple domains, and **very large training sets were adopted** (CUHK03)

- Zhang et al. [24] it learns a classifier specifically for each person (this model characteristic **can benefit when large training sets are employed**)

	VIPeR	PRID450s	CUHK01
Images	1264	900	3884
Individuals (ID)	632	450	971
Images per ID (per view)	1	1	2

[11] T. Xiao, H. Li, W. Ouyang, and X. Wang, “Learning deep feature representations with domain guided dropout for person re-identification,” in CVPR, 2016.

[24] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, “Sample-specific svm learning for person re-identification,” in CVPR, 2016.

Case 1:
State-of-the-art comparison

Rank	1	5	10	20
CUHK01				
FT-JSTL+DGD[11]	**66.60**	-	-	-
LSSCDL [24]	65.97	≈ **88.0**	≈ **92.0**	≈ 96.0
Our	59.63	83.66	89.71	94.39
Deep+LOMO [9]	55.51	78.40	83.68	92.59
3TCP [4][8]	53.70	84.30	91.00	**96.30**
CMC [13]	53.40	76.40	84.40	90.50
Mirror [10]	40.40	64.63	75.34	84.08

- Xiao et al. [11] was designed to learn features from multiple domains, and **very large training sets were adopted** (CUHK03)
- Zhang et al. [24] it learns a classifier specifically for each person (this model characteristic **can benefit when large training sets are employed**)

Rank	1	
VIPeR		
Our	**58.77**	**8[illegible]**
LSSCDL [24]	42.66	
FT-JSTL+DGD[11]	38.60	

	VIPeR	PRID450s	CUHK01
Images	1264	900	3884
Individuals (ID)	632	450	971
Images per ID (per view)	1	1	2

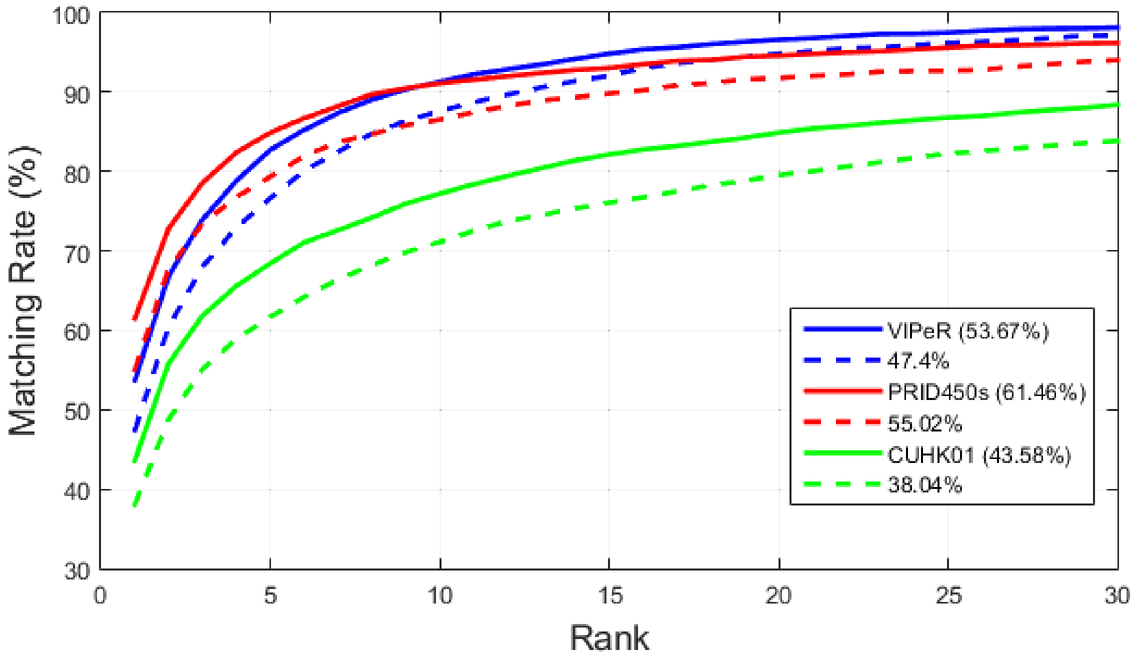
Rank	1	5
PRID450s		
Our	**71.56**	**90.[illegible]**
Deep+LOMO [9]	66.62	86.[illegible]
LSSCDL [24]	60.49	-

[11] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in CVPR, 2016.

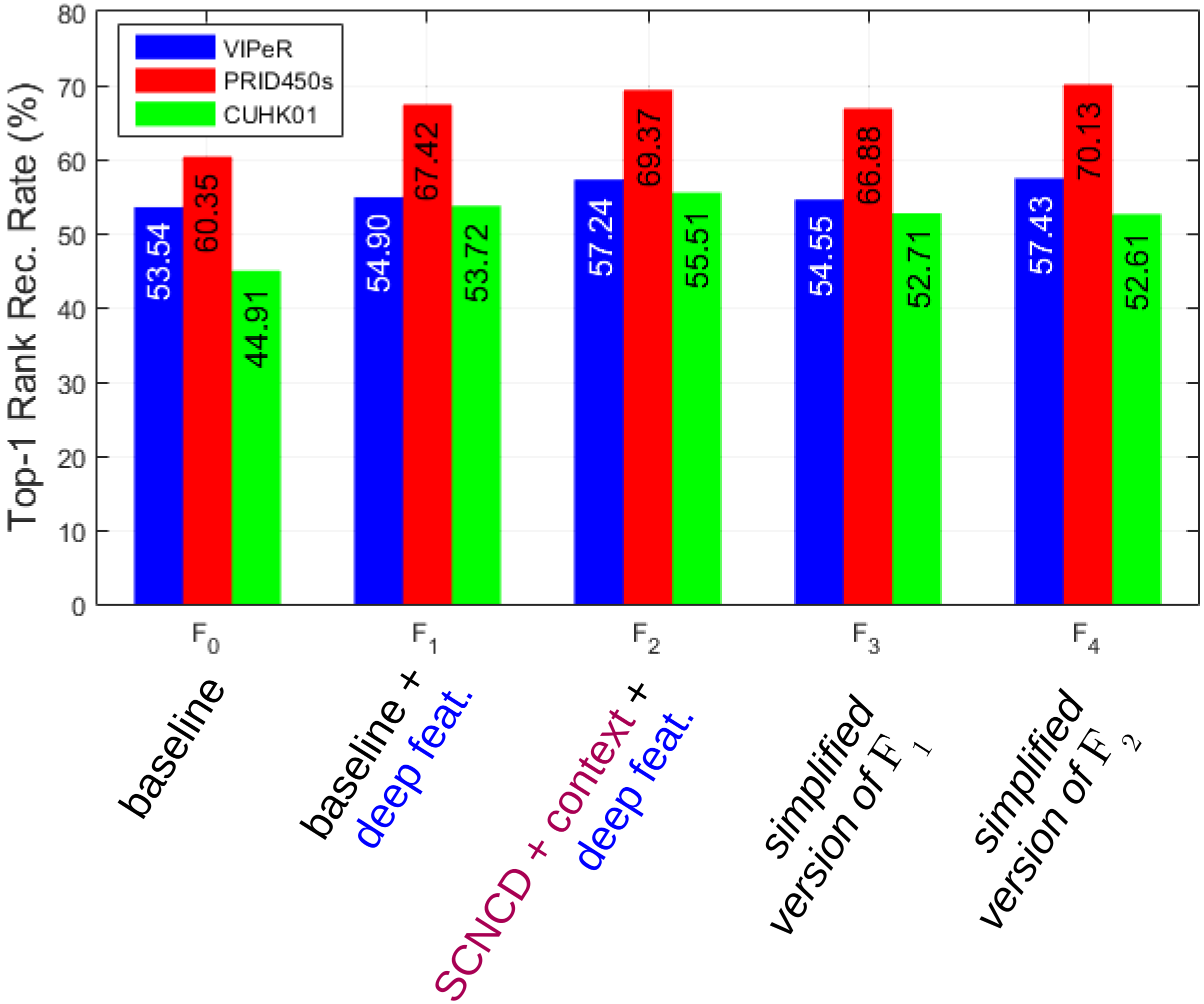
[24] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-specific svm learning for person re-identification," in CVPR, 2016.

Case 2: Context information

- Used just C_5 and C_6 ,with (solid line) and without context information
- It **improved the overall accuracy** on three evaluated datasets
- Different from Yang et al. [6], we evaluated SCNCD using different color models, a more powerful strategy *(DDN)* [21] and different similarity function *(Polynomial feat. map)* [7].



Case 3: Complementary features



- All complementary features **outperformed the baseline**
- **Simplifications** still have **strong discriminative power** and require less computation resources.
- The **benefit of deep feature** can be seen when we compare F_0 with F_1
- F_2 obtained **best overall accuracy**.

  Despite extraction procedures, it is more compact than F_1

Conclusions

- We exploited **different feature representations**, combined with a ranking aggregation strategy to advance re-id.
  - The proposed new features demonstrated to **complement each other**, being very powerful when combined with a ranking aggregation strategy.
- We show that **hand-crafted and deep features fusion** can improve re-identification performance especially in domains where there is a reduced amount of available data.

Exploiting feature representations through similarity learning and ranking aggregation for person re-identification

Julio C. S. Jacques Junior [1,2]
juliojj@gmail.com

Xavier Baró [2,3]
Sergio Escalera [1,2]

[1] *Department of Mathematics and Informatics* - University of Barcelona, Spain
[2] *Computer Vision Center* - Universitat Autònoma de Barcelona, Spain
[3] *Faculty of Computer Science, Multimedia and Telecommunication* - Universitat Oberta de Catalunya, Spain

Computational cost *

Task	Time (in seconds)		
Extract contextual information (per image)	0.146		
SCNCD feature extraction (per image)	0.131		
Deep feature extraction per image (provided in [9]) **	1.0		
Baseline feature extraction (per image)	0.069		
Feat. Representation (per image) → Build C_1 to C_4	5.348		
Feat. Representation (per image) → Build C_5 and C_6	0.063		
Apply PCA on C_7 (per image)	0.063		
Learning stage (single run on the whole VIPeR dataset) **Test** stage (each probe image on VIPeR)	F_1 = 239.7	|	Test = 0.014
	F_2 = 125.8	|	Test = 0.007
	F_3 = 194.2	|	Test = 0.011
	F_4 = 102.8	|	Test = 0.006
Ranking aggregation (per image)	0.1473		

* Adapted MATLAB implementation from [5], using a 2.30Hz Intel Core i7 CPU and 8Gb of memory.

** Using a 2.00Hz Xeon CPU with 16 cores.