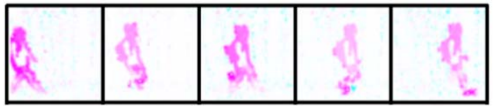
Deep multi-task learning for gait-based biometrics

- **Manuel** J. Marín-Jiménez (*Univ. of Córdoba*)
- Francisco M. Castro (*Univ. of Málaga*)
- Nicolás Guil (*Univ. of Málaga*)
- Fernando de la Torre (*Carnegie Mellon University*)
- Rafael Medina-Carnicer (*Univ. of Córdoba*)

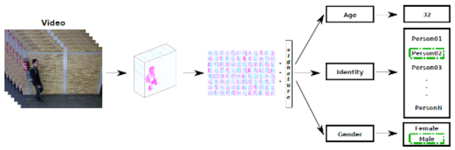
Outline

1. **Problem definition**

2. Our approach
   i. Input data



   ii. Deep Multi-task Model



3. Experiments and results

4. Conclusions and future work

Motivation

- *“Who’s he?”*
- *“I cannot distinguish his face”*
- *“But he **walks** like Peter”*

We are good at identifying people at a **distance**

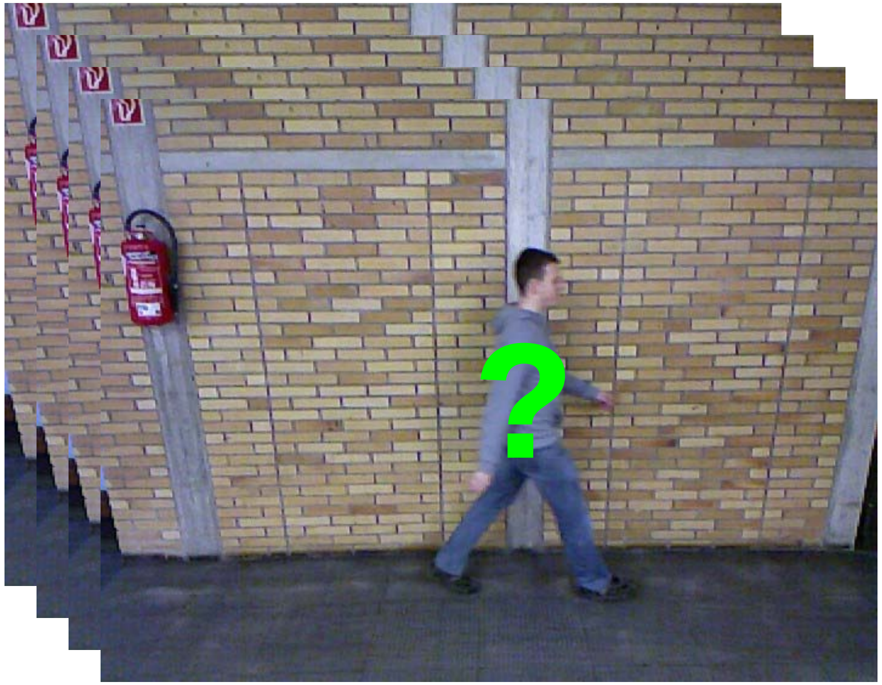
- But, why?
- Because each person has his/her own gait pattern → **gait signature**

The problem

Objective: identify people based on **the way they walk**

→ **Gait recognition**

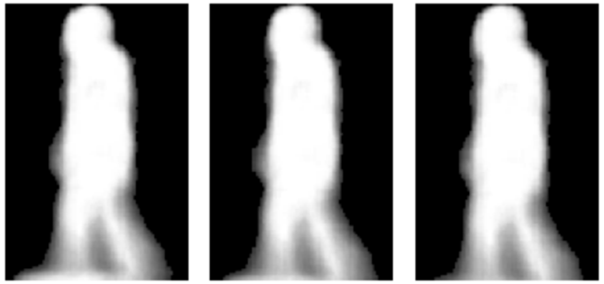
+ Gender
+ Age



Input: video sequence
Output: identity

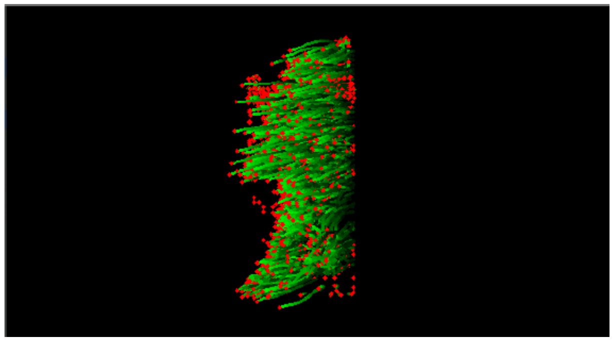
Previous approaches

Context: video surveillance, control access,...



Gait Energy Image (GEI)
[Han PAMI06]

Chrono-Gait Image (CGI)
[Wang PAMI12]



Pyramidal Fisher Motion (PFM)
[Castro IJPRAI17]

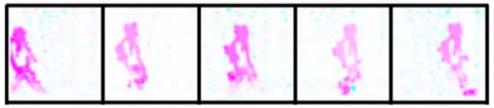
Hand-crafted!

*Images extracted from their corresponding papers: [Han PAMI06], [Wang PAMI12], [Castro IJPRAI17]

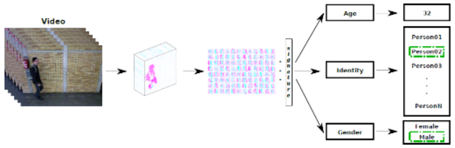
Outline

1. Problem definition

2. **Our approach**
   i. Input data



   ii. Deep Multi-task Model

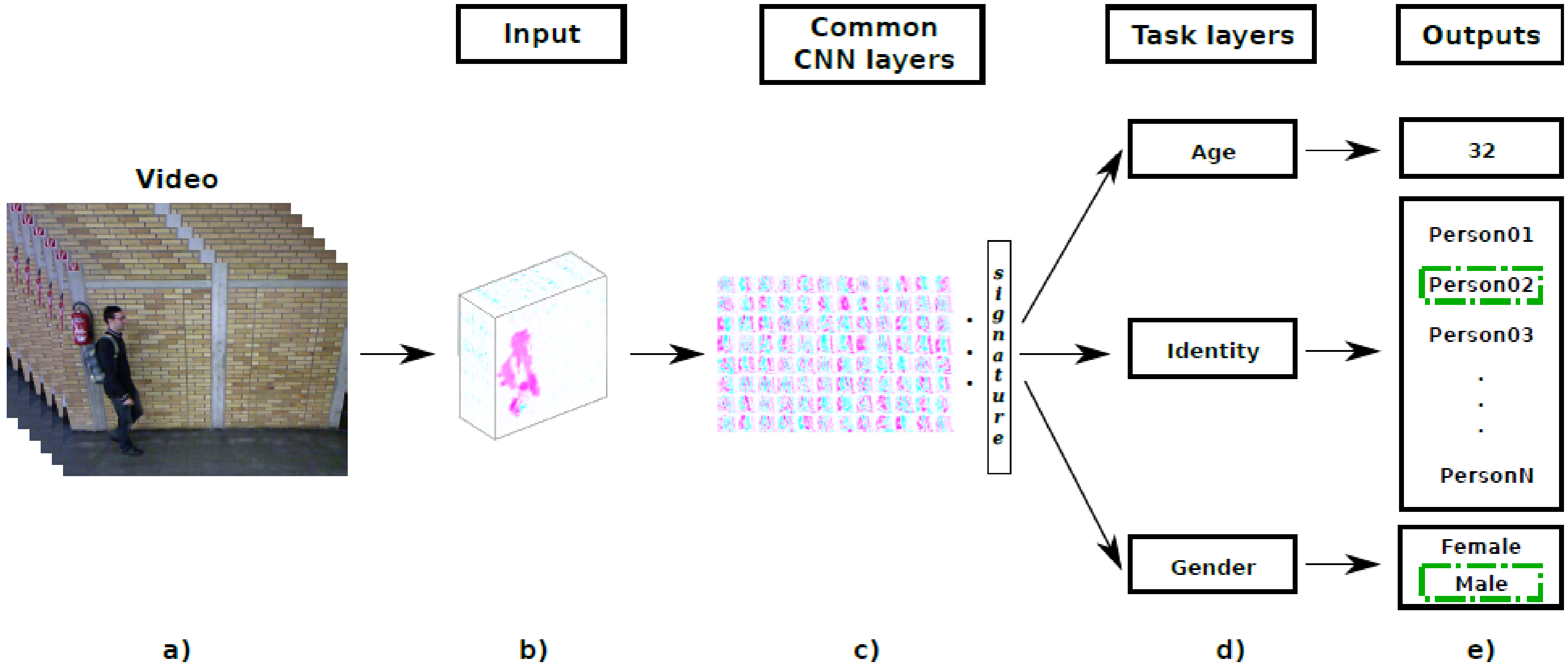


3. Experiments and results

4. Conclusions and future work

Multi-task CNN

- Goal: identify people + age + gender

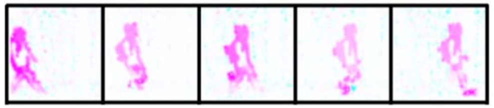


Outline

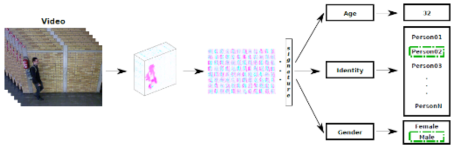
1. Problem definition

2. Our approach

   i. **Input data**



   ii. Deep Multi-task Model

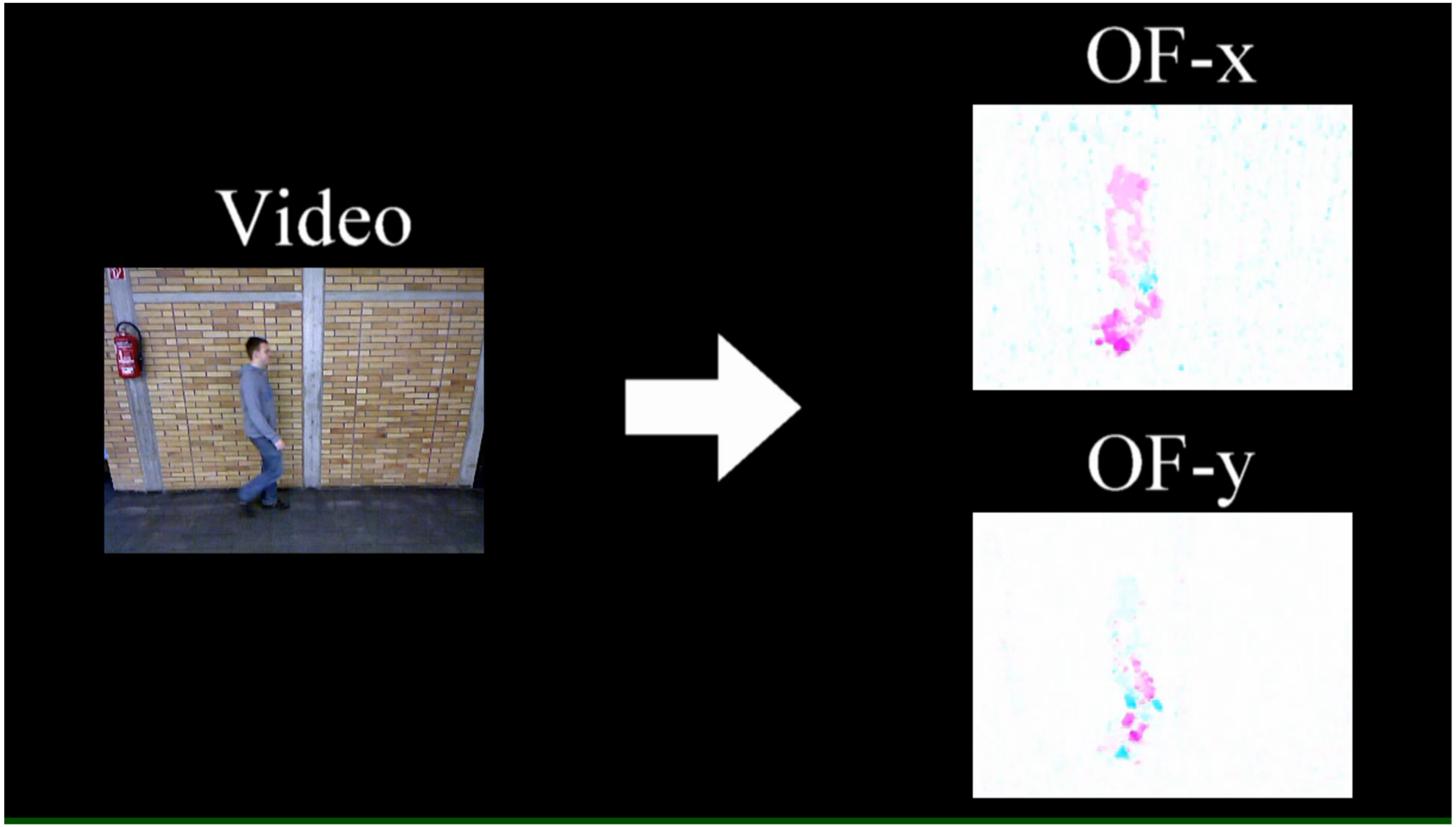


3. Experiments and results

4. Conclusions and future work

Input data

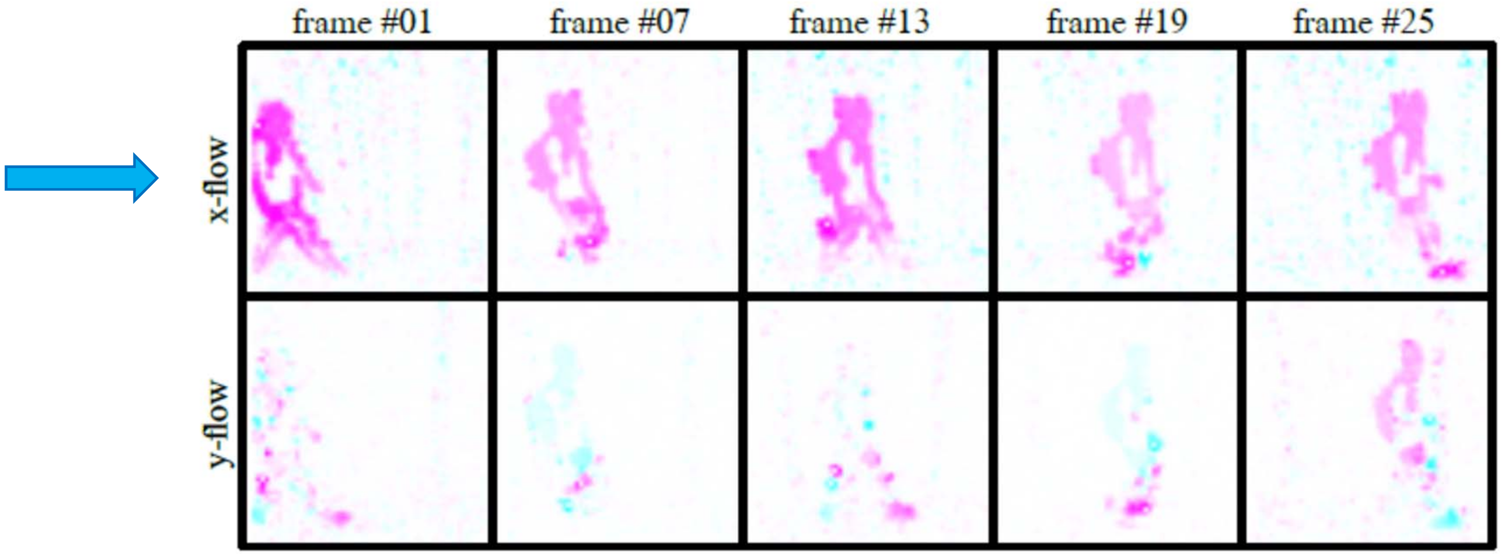
- Optical flow channels: {OF-*x*, OF-*y*} @ 80x60 pix



[Castro17] FM Castro, MJ Marin-Jimenez, N. Guil and N. Perez de la Blanca, “Automatic learning of gait signatures for people identification” in IWANN, 2017

Input data

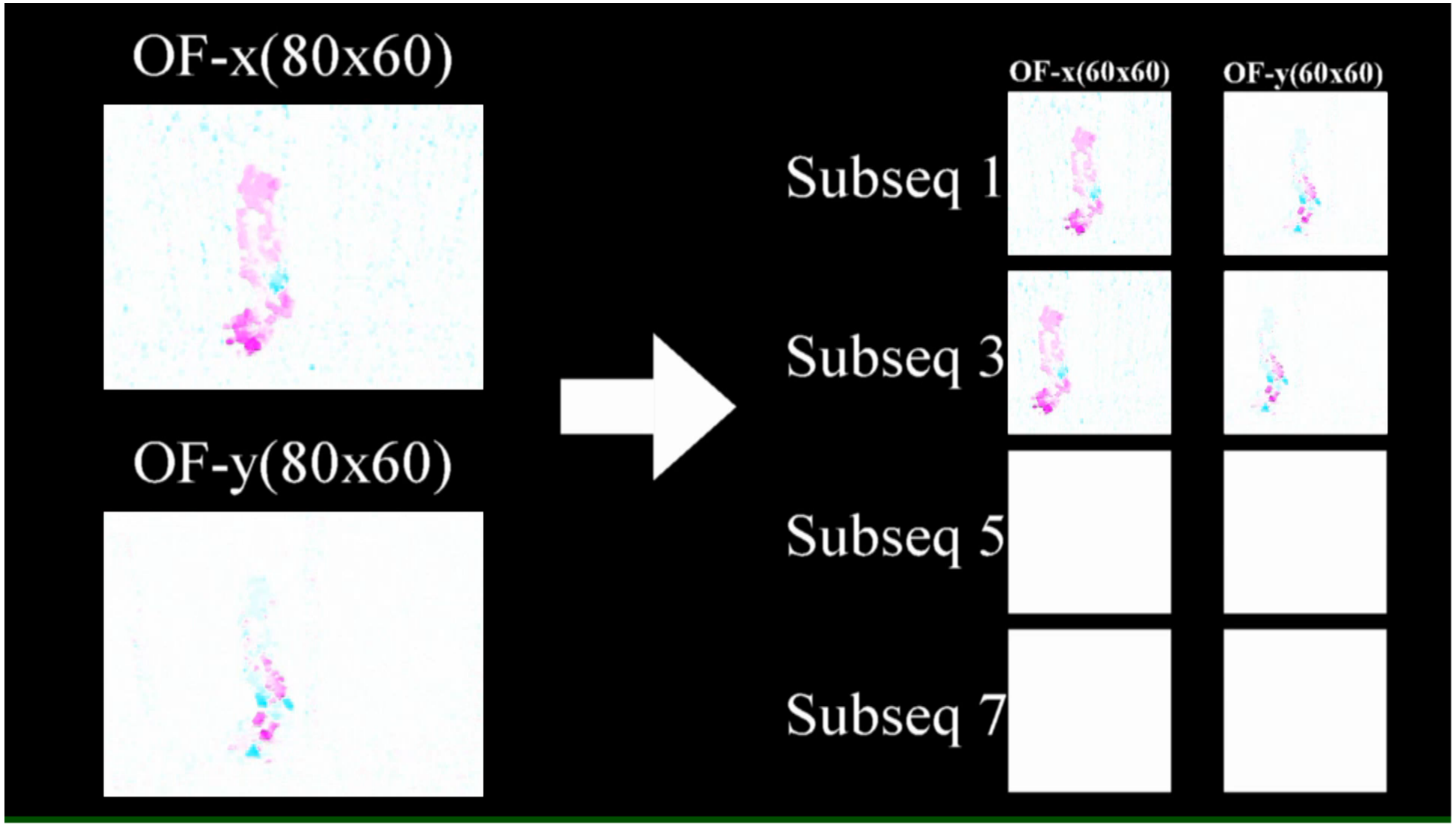
- Optical flow channels: {OF-*x*, OF-*y*}
- Fixed length: 25 frames (~1 gait cycle)
- Crop frames to 1:1 aspect ratio
- Person centred in middle-frame



[Castro17] FM Castro, MJ Marin-Jimenez, N. Guil and N. Perez de la Blanca, "Automatic learning of gait signatures for people identification" in IWANN, 2017

Input data

- Multiple subsequences are extracted

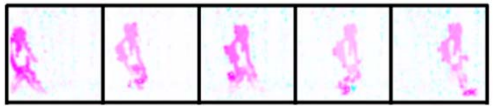


[Castro17] FM Castro, MJ Marin-Jimenez, N. Guil and N. Perez de la Blanca, “Automatic learning of gait signatures for people identification” in IWANN, 2017

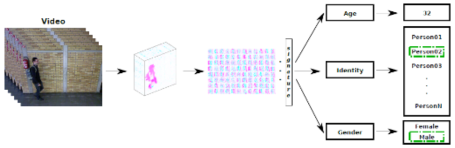
Outline

1. Problem definition

2. Our approach
   i. Input data



   ii. Deep Multi-task Model



3. Experiments and results

4. Conclusions and future work

Deep Multi-task (DMT) model

- **DMT loss** function:

  identification loss (main) + auxiliary losses.

CNN filters

$$\mathcal{L}_{\mathrm{DMT}}(g(\mathbf{v},\theta),\mathbf{Y}) = \mathcal{L}_m(\hat{\mathbf{y}}^m, y^m) + \sum_{t=1}^{T} \lambda_t \cdot \mathcal{L}_t(\hat{\mathbf{y}}^t, y^t)$$

ground-truth

CNN output

weight task *t*

Main task: identification

- Identification loss function:

  softmax log-loss

$$\mathcal{L}_m(\hat{\mathbf{y}}, c) = -\hat{y}_c + \log \sum_{k=1}^{C} e^{\hat{y}_k}$$

ground-truth

c-th component

CNN output

0	0	0	0	0	0	0	1	0	0

Aux task: gender recognition

- Gender loss function:

  softmax log-loss (two classes)

$$\mathcal{L}_g\left(\hat{\mathbf{y}}, c\right) = -\hat{y}_c + \log \sum_{k=1}^{C} e^{\hat{y}_k}$$

ground-truth

CNN output

c-th component

0	1

Aux task: age estimation

- Age loss function:

  Tukey's biweight loss [Black96] → regression

$$\mathcal{L}_a(\hat{y}_i, y_i) = \rho(r_i^{\text{MAD}})$$

$c = 4.6851$

$$\rho(r_i) = \begin{cases} \frac{c^2}{6}[1 - (1 - (\frac{r_i}{c})^2)^3] & , \quad \text{if} |r_i| \leq c \\ \frac{c^2}{6} & , \quad \text{otherwise} \end{cases}$$

$$r_i^{\text{MAD}} = \frac{y_i - \hat{y}_i}{1.4826 \times \text{MAD}}$$

residual

$$\text{MAD} = \underset{k \in \{1,\ldots,S\}}{\text{median}} \left(\left| r_k - \underset{j \in \{1,\ldots,S\}}{\text{median}} (r_j) \right| \right)$$

[Black96] Michael J Black and Anand Rangarajan, "On the unification of line processes, outlier rejection, and robust statistics with applications in early vision", IJCV, vol. 19, no. 1, pp. 57–91, 1996

Aux task: identity verification

- Verification loss function:

  L2 distance with margin [Hadsell06]

$$V(\mathbf{f}_i,\mathbf{f}_j,y_{ij})=\begin{cases} \frac{1}{2}||\mathbf{f}_i-\mathbf{f}_j||_2^2, & \text{if } y_{ij}=1 \\ \frac{1}{2}\max(0,m-||\mathbf{f}_i-\mathbf{f}_j||_2)^2, & \text{if } y_{ij}=-1 \end{cases}$$

feature vectors
(last FC layer)

margin

+1: same id
-1: different id

[Hadsell06] Raia Hadsell, Sumit Chopra, and Yann LeCun, "Dimensionality reduction by learning an invariant mapping," in CVPR, 2006, vol. 2, pp. 1735–1742

CNN architectures

Convolutional: spatial filters (**trainable**, local) + **ReLU**

Fully-connected

Softmax: values in (0,1) and adds up to 1

CNN type **A**

conv1	conv2	conv3	conv4	full5	softmax
7x7x64 stride 1 pool 2x2	5x5x128 stride 2 pool 2x2	3x3x512 stride 1 pool 2x2	2x2x512 stride 1	256	N units (classes)

CNN type **B**

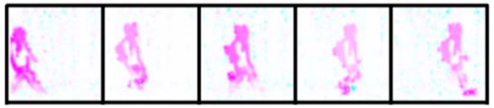
conv1	conv2	conv3	conv4	full5	softmax
7x7x64 stride 1 norm pool 2x2	5x5x128 stride 2 pool 2x2	3x3x512 stride 1 pool 2x2	2x2x2048 stride 1	1024 dropout 0.1	N units (classes)

Gait signature

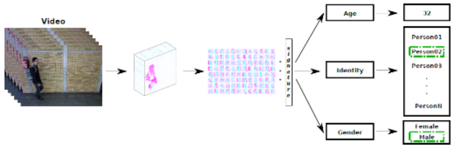
Outline

1. Problem definition

2. Our approach
   i. Input data



   ii. Deep Multi-task Model



3. Experiments and results

4. Conclusions and future work

Dataset

TUM-GAID dataset [Hofmann JVCIR14]

- 305 subjects (100 train + 50 val + **155 test**)
- Scenarios:
  - Normal (N) + elapsed time (**TN**)
  - Bag (B) + elapsed time (**TB**)
  - Coating shoes (S) + elapsed time (**TS**)
- **Identity, age** and **gender** labels

[Hofmann JVCIR14] M. Hofmann et al. The TUM Gait from Audio, Image and Depth (GAID) database: multimodal recognition of subjects and traits. J. of Visual Com. and Image Repres. 2014

Implementation details

- DMT training:
  - *task-wise early stopping* criterion [Zhang14]
  - back-propagation → SGD+momentum
  - batch: 256 samples
  - learning rate: 0.01 (reduced 0.1 if val error stuck)
  - max epochs: 30

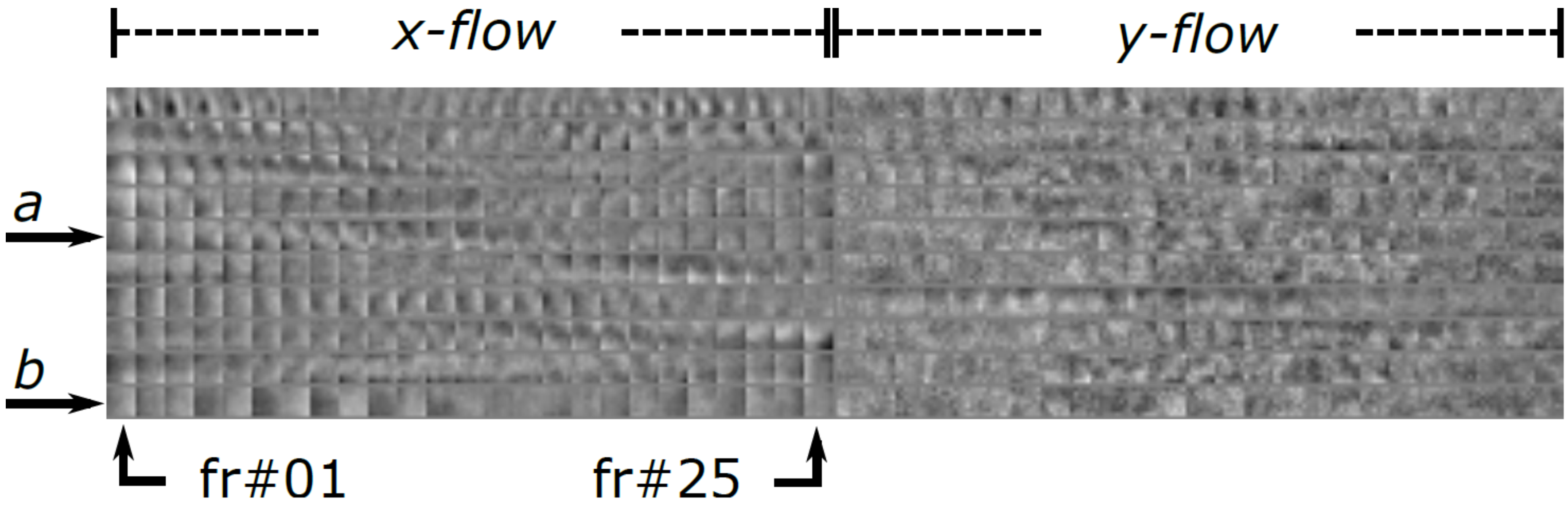
- From subseqs to sequence-level decision:
  - Majority voting (e.g. on SVM scores)
  - Product of Softmax scores

[Zhang14] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Facial landmark detection by deep multitask learning," in ECCV, 2014, pp. 94–108

Filters learnt

Convolutional filters @ first layer

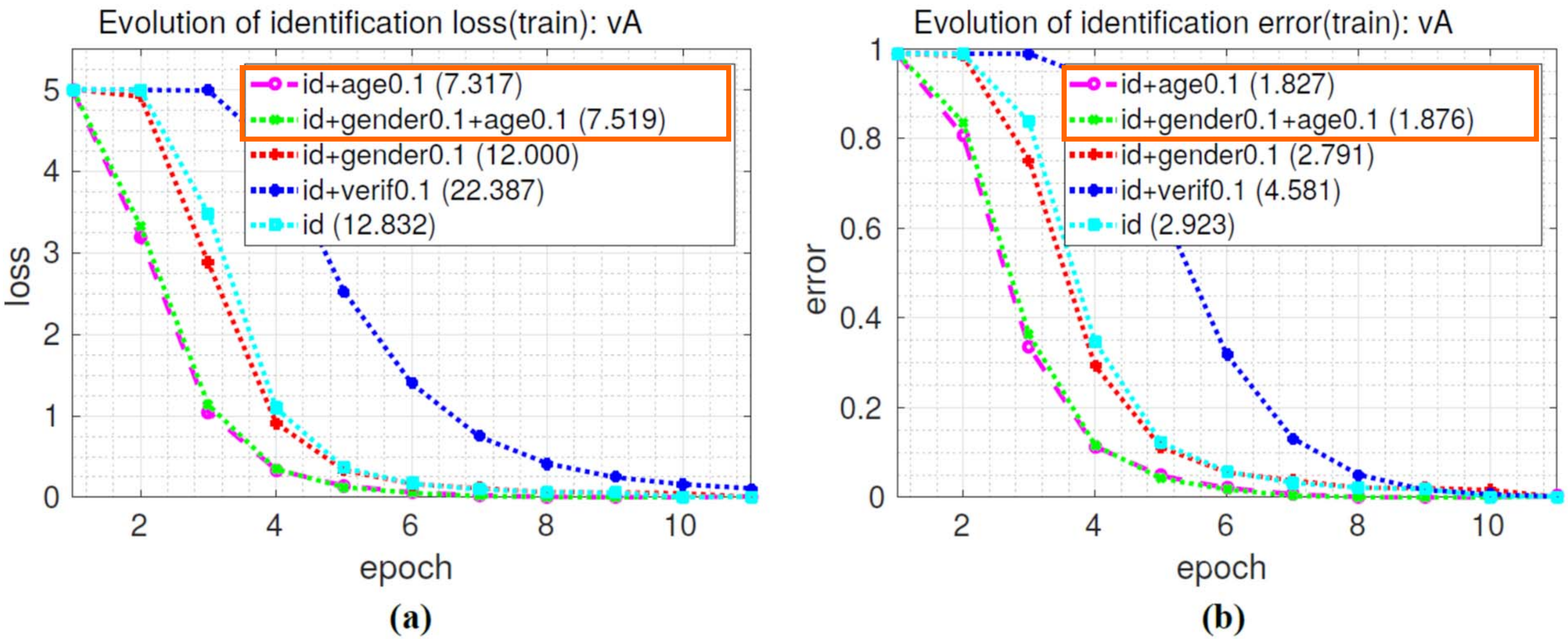
- *a*: **spatial** derivatives
- *b*: **temporal** derivatives



@mjmarinj

Experiment: aux tasks contrib.

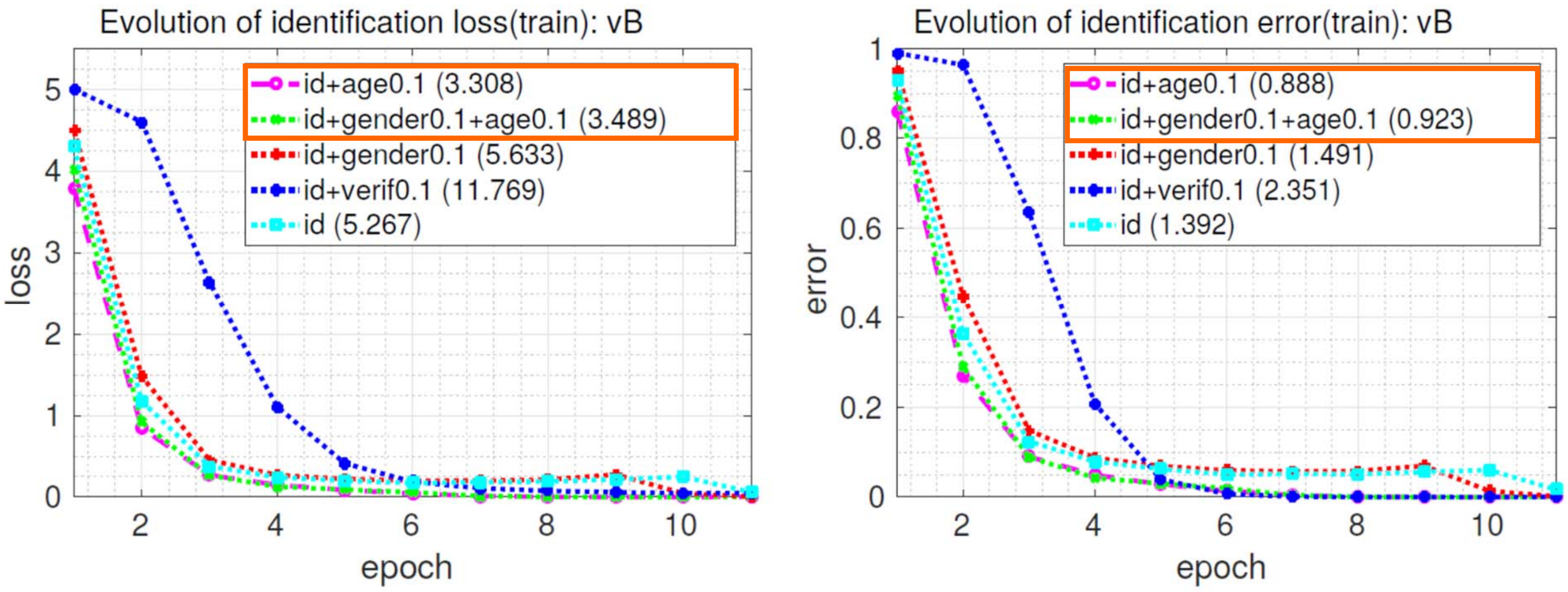
- Auxiliary tasks **speed up convergence** of the main task.



AUC in parenthesis: lower is better

Experiment: aux tasks contrib.

- Auxiliary tasks **speed up convergence** of the main task.

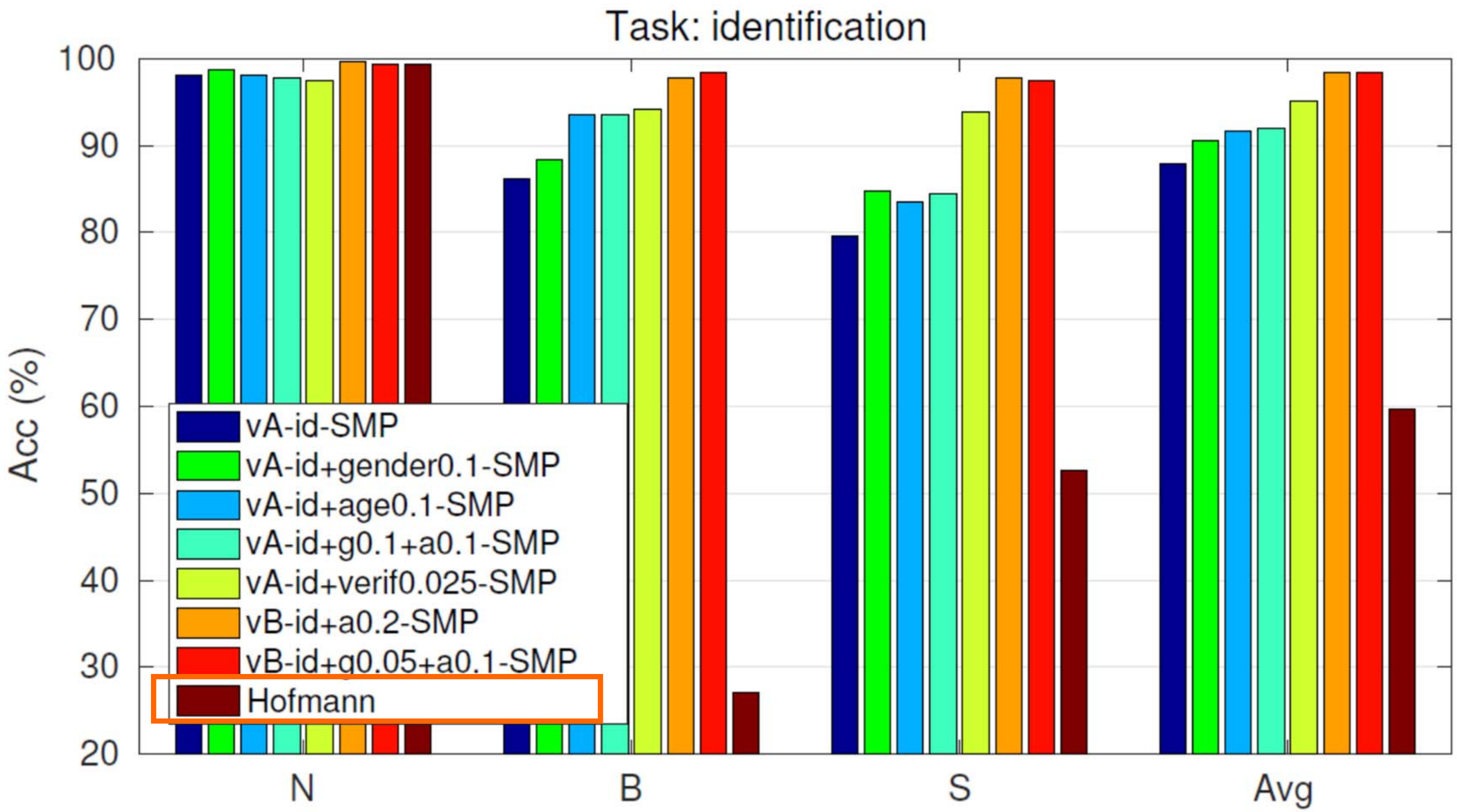


(c) (d)

AUC in parenthesis: lower is better

Experiment: identification

- Identification results



SMP: softmax product

Accuracy:
higher is better

State-of-the-art: comparison

CNN type **B**

conv1	conv2	conv3	conv4	full5	softmax
7x7x64 stride 1 norm pool 2x2	5x5x128 stride 2 pool 2x2	3x3x512 stride 1 pool 2x2	2x2x2048 stride 1	1024 dropout 0.1	N units (classes)

Castro et al. IWANN2017

conv1	conv2	conv3	conv4	full5	full6	softmax
7x7x96 stride 1 norm pool 2x2	5x5x192 stride 2 pool 2x2	3x3x512 stride 1 pool 2x2	2x2x4096 stride 1	4096 dropout	2048 dropout	

Experiment: identification

	Method	N	B	S	*Avg*	TN	TB	TS	*Avg*
80x60	Ours-1 (SVM)	100	97.1	97.1	98.1	53.1	59.4	50	54.2
	Ours-2 (SVM)	99.7	96.5	97.4	97.9	56.3	56.3	56.3	56.3
	Ours-2 (7-NN)	99.7	97.4	99.7	**98.9**	59.4	62.5	68.8	**63.6**
	Castro17b-CNN (SVM)	99.7	97.1	97.1	98	59.4	50	62.5	57.3
640x480	Hofmann et al	99.4	27.1	52.6	59.7	44	6	9	19.7
	RSM	100	79	97	92	58	38	57	51.3
	Castro17a-PFM	99.7	99	99	**99.2**	78.1	56.3	46.9	**60.4**

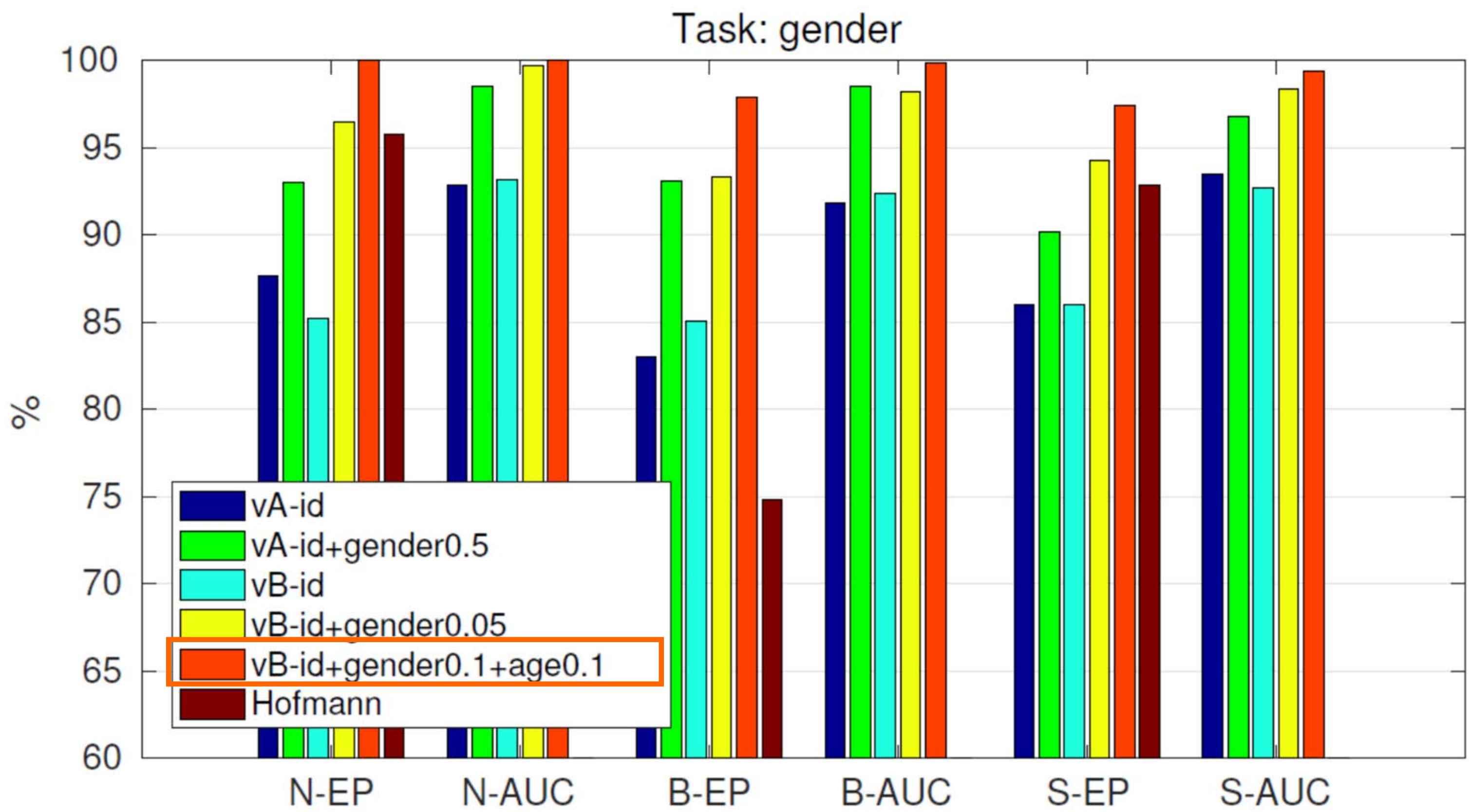
Ours-1: id+age0.2 (vB)
Ours-2: id+verif0.1 (vB)
7-NN: 7-Nearest Neighbour with PCA-128

[Castro17a] FM Castro, MJ Marín-Jiménez, N. Guil, R. Muñoz-Salinas, "Fisher Motion Descriptor for Multiview Gait Recognition" in IJPRAI 31(1): 1-40, 2017

[Castro17b] FM Castro, MJ Marin-Jimenez, N. Guil and N. Perez de la Blanca, "Automatic learning of gait signatures for people identification" in IWANN, 2017

Experiment: gender

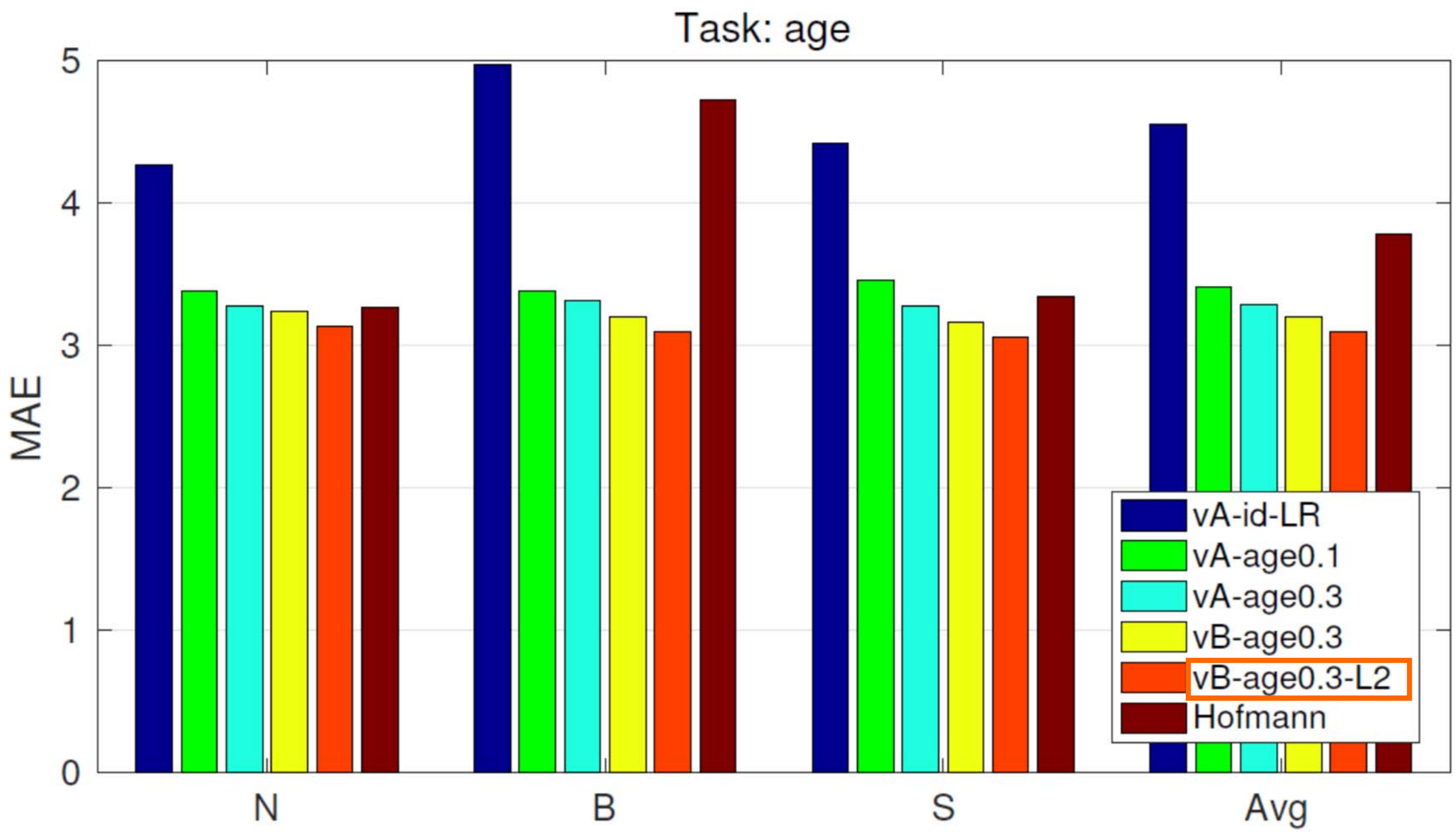
- Gender recognition results



Accuracy:
higher is better

Experiment: age

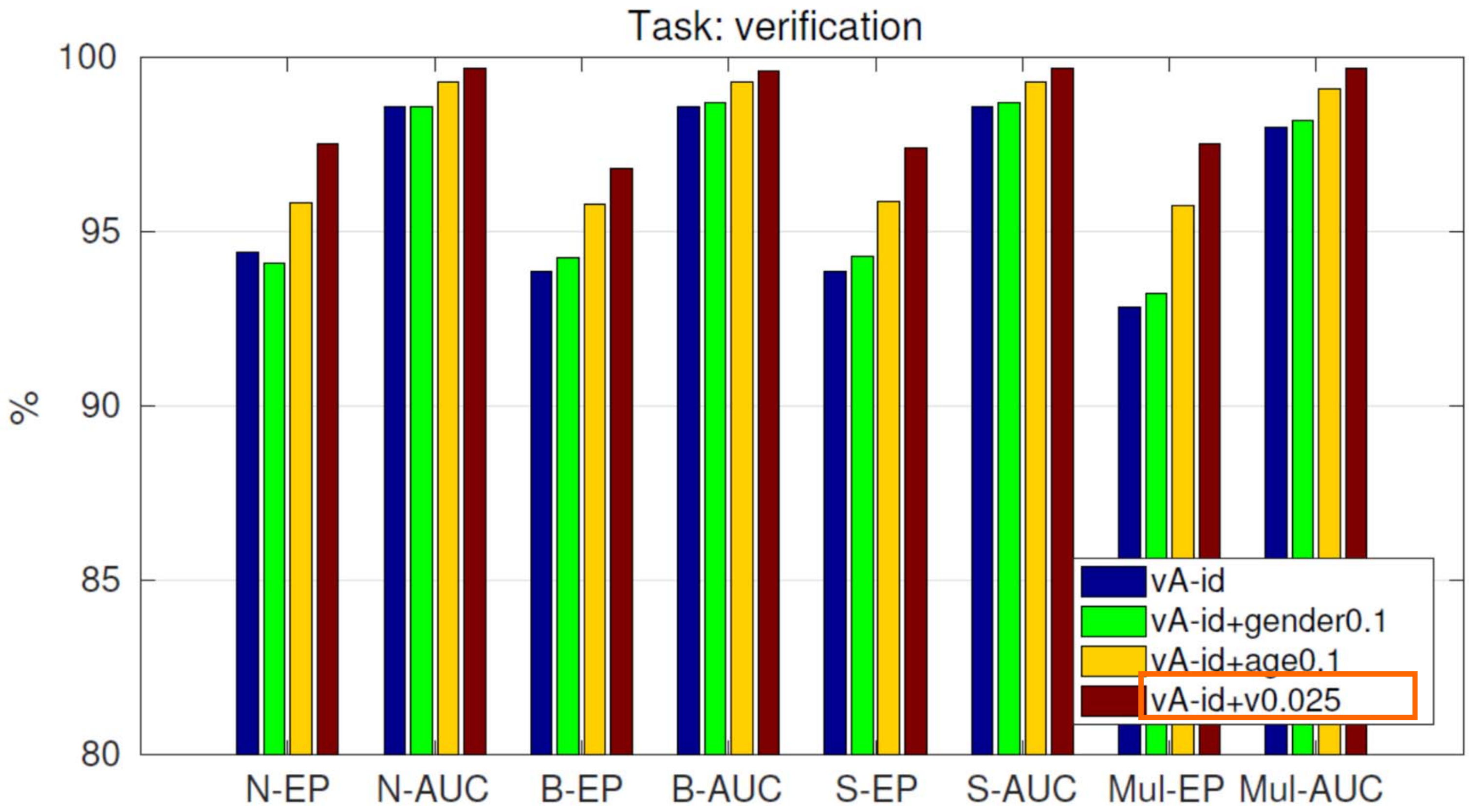
- Age estimation results



Mean Absolute Error: lower is better

Experiment: verification

- Identity verification results



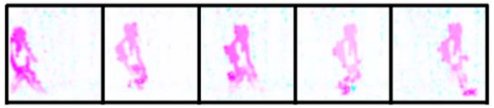
EP: accuracy at equilibrium point

AUC: area under the precision-recall curve

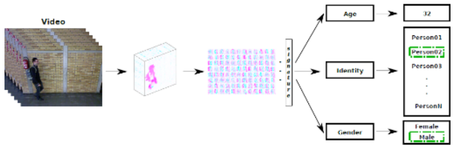
Outline

1. Problem definition

2. Our approach
   i. Input data



   ii. Deep Multi-task Model



3. Experiments and results

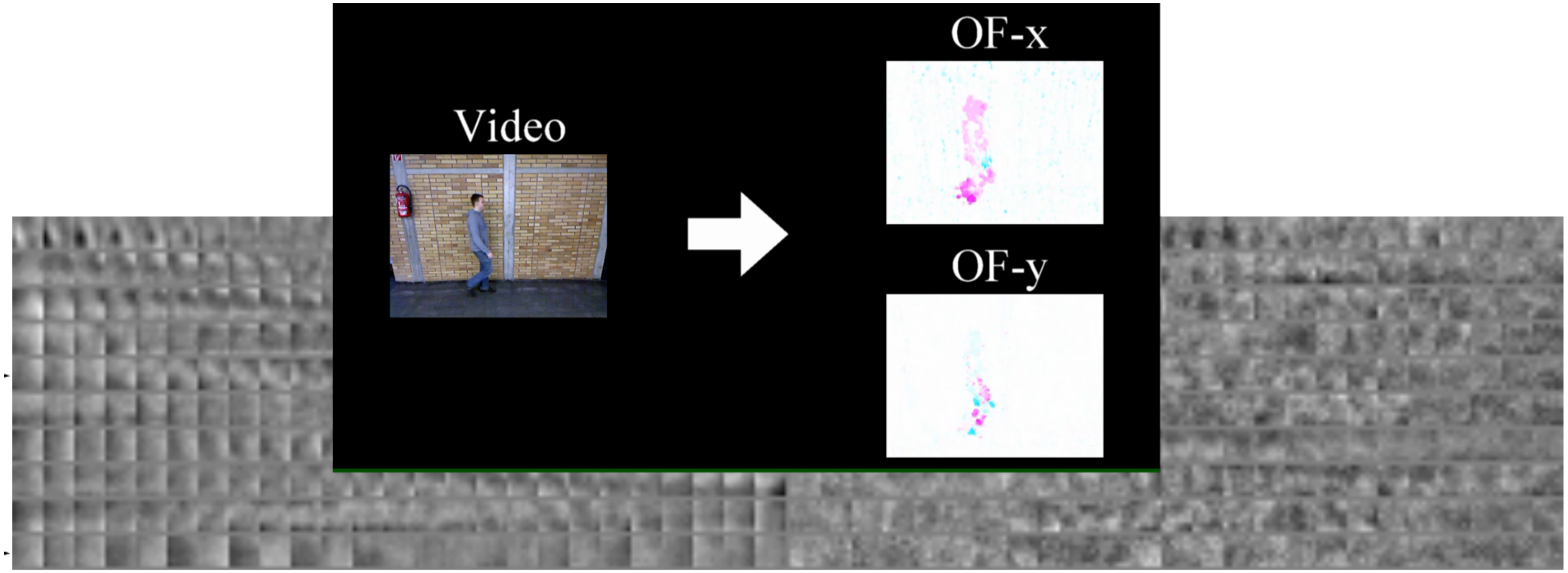
4. **Conclusions and future work**

Conclusions and future work

- DMT speeds up convergence of the main task.
- Accuracy of identification (main task) improves.
- CNN filters of the first layers are useful for several tasks.
- Other modalities: gray, depth,...
- Other tasks

CNN gait

THANK YOU
FOR YOUR ATTENTION



Manuel Jesús Marín Jiménez *@mjmarinj*