DEEP MULTI-TASK LEARNING FOR GAIT-BASED BIOMETRICS

M.J. Marín-Jiménez[1], *F.M. Castro*[2], *N. Guil*[2], *F. de la Torre*[3], *R. Medina-Carnicer*[1]

[1]University of Cordoba; [2]University of Malaga; [3]Carnegie Mellon University

ABSTRACT

The task of identifying people by the way they walk is known as '*gait recognition*'. Although gait is mainly used for identification, additional tasks as gender recognition or age estimation may be addressed based on gait as well. In such cases, traditional approaches consider those tasks as independent ones, defining separated task-specific features and models for them. This paper shows that by training jointly more than one gait-based tasks, the identification task converges faster than when it is trained independently, and the recognition performance of multi-task models is equal or superior to more complex single-task ones. Our model is a multi-task CNN that receives as input a fixed-length sequence of optical flow channels and outputs several biometric features (identity, gender and age).

Index Terms— Multi-task learning, Deep Neural Networks, Gait Recognition, Biometrics

1. INTRODUCTION

This paper addresses the problem of multi-task learning [1] for automatic extraction of multipurpose gait signatures (*i.e.*descriptors) for gait recognition (to identify people by the way the walk). Gait is considered a non-invasive type of biometrics, as it does not need the cooperation of the subject and can be performed at a distance, in contrast to other biometric approaches as fingerprint or iris recognition. It is very challenging as the differences between gait styles of different people can be very subtle, and the gait style of the same person can be altered by factors as the type of floor, footwear, clothing or carried objects (*e.g.*bag). Most gait approaches found in the literature rely on binary silhouettes [2] obtained by background segmentation, which discard information that may be crucial in particular situations. However, state-of-the-art results are obtained when motion information is directly extracted from the RGB video source [3]. Therefore, we will build our approach on motion information not extracted from silhouettes but from the original pixels.

In recent years, deep learning (DL) approaches [4] have shown that state-of-the-art results can be achieved in different computer vision problems by providing enough training data to purpose-specific deep neural networks. However, only a few works can be found in the literature [5, 6, 7] where DL approaches have been applied to gait recognition. In our opinion, the main reason is the fact that the relative amount of training data for gait is substantially smaller than for other computer vision problems as object recognition [8], and DL is highly data-demanding. On the other hand, although gait is mainly used for people identification (*i.e.*selecting one subject from a gallery set), other biometric tasks can be defined based on gait, *e.g.*gender recognition or age estimation [9, 10]. However, not much attention has been paid to the fact that those tasks are closely related and can benefit ones from others when are jointly considered. We tackle in this work the following research question: *can we improve the training speed and accuracy of a given gait-based task by using a multi-task learning approach?*. In this work we automatically learn gait signatures from volumes of optical flow channels, instead of hand-crafted descriptors of binary silhouettes, by using a multi-task convolutional neural network (CNN), obtaining gait signatures that are simultaneously suitable for the tasks of identification, gender recognition, age estimation and identity verification. Fig. 1 summarizes our key idea.

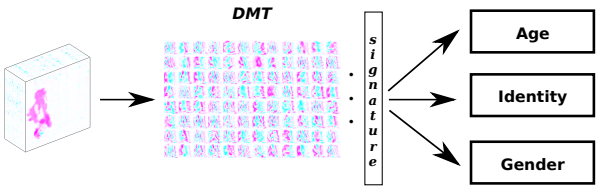

Fig. 1: **Pipeline for gait-based multi-task learning**. The first layers of the convolutional network are shared among the different tasks. Then, task-specific layers are added.

The main contribution of this paper is the definition of a multi-task CNN for the problem of gait-based biometrics, along with its corresponding loss functions for training it. A thorough experimental study on the challenging TUM-GAID dataset supports the main research hypothesis of this work: (*i*) multi-task learning speeds up the velocity of convergence of the objective function of the main task (*i.e.*identification), and (*ii*) improves the capacity of generalization of the shared gait signatures obtained for the biometric tasks.

The rest of the paper is organized as follows. In Sec. 2 we describe our deep multi-task (DMT) model. Then, Sec. 3 contains the experiments and results. Finally, the conclusions and future work are presented in Sec. 4.

This work has been funded under projects TIC-1692 (Junta de Andalucía) and TIN2016-75279-P (Spanish Ministry of Science and Tech.). The GPU Titan X Pascal used for this research was donated by NVIDIA.

2. DEEP MULTI-TASK MODEL

Our goal is to learn a deep model which is able to automatically extract discriminant features from gait sequences that can be simultaneously applied to several tasks: identification, gender recognition, age estimation and verification. We will use the following nomenclature along this paper, where vectors and matrices are marked in bold-face: $\mathbf{V}$: RGB video input; $\mathbf{v}_i$: preprocessed video clip i, *e.g.*optical flow volume; y_i^t: label for sample $\mathbf{v}_i$ and task t (superscript t can be omitted for a given task); $g(\mathbf{v}, \theta)$: non-linear function applied to $\mathbf{v}$ with a set of parameters θ, including the trained CNN filters; $\hat{\mathbf{y}}_i$: output of the network for input $\mathbf{v}_i$, *i.e.*$\hat{\mathbf{y}}_i = g(\mathbf{v}_i, \theta)$; and, $\mathbf{f}_i^l$: feature vector extracted from sample $\mathbf{v}_i$ at layer l, *i.e.*output of fully-connected layer l.

To train a deep multi-task (DMT) model with $T + 1$ tasks, we need a set of S tuples of the following form: $(\mathbf{v}_i, y_i^m, y_i^1, y_i^2, \ldots, y_i^T)$, where y_i^m corresponds to the main task and the remaining y_i^t to the auxiliary tasks.

The multi-task loss function $\mathcal{L}_{\text{DMT}}$ for a given sample $\mathbf{v}$ is defined by the following equation:

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

where $\mathbf{Y} = (y^m, y^1, y^2, \ldots, y^T)$, $\mathcal{L}_m$ is the loss function of the main task and λ_t is the weight given to the auxiliary task t with corresponding loss $\mathcal{L}_t$.

2.1. Single-task CNN architecture

First of all, we define the shape of the input data for our CNN. Since the video sequences $\mathbf{V}$ may have different length, we extract overlapping clips $\mathbf{v}_i$ of 25 frames, following the criterion of [11]. After resizing the video frames to a common size of 80×60 pixels, as suggested in [11], the input data of the CNN are cuboids of optical flow [12] channels (*i.e.*horizontal and vertical components). The set of frames is cropped (*i.e.*removing columns) in such a way that the person in the middle frame of the volume is approximately x-centered, thanks to a coarse detection of the person by background subtraction [13]. In particular, each cuboid has size: $60 \times 60 \times 50$.

Our CNN for gait signature extraction is compound of four convolutional layers followed by two fully-connected layers. Two versions of different complexity (*i.e.*number of intermediate filters) are evaluated in this work: *vA* and *vB*. The baseline configuration of the CNN architectures is summarized in Tab. 1. Note that both networks are simplified versions of the one used in [11], that we have found suitable for the purpose of this paper. All convolutional layers, which include max-pooling, are followed by a Rectified Linear Unit (ReLU) layer. At training time, the corresponding loss layers that will be defined below are added on top of 'Full02'. Whereas, at deployment time, a Softmax layer is added on top of 'Full02' for the classification-like tasks.

V	Conv01	Conv02	Conv03	Conv04	Full01	Full02
vA	$7 \times 7 \times 64$ P: 2×2	$5 \times 5 \times 128$ P: 2×2	$3 \times 3 \times 512$ P: 2×2	$2 \times 2 \times 512$	256	C
vB	$7 \times 7 \times 64$ P: 2×2; Norm	$5 \times 5 \times 128$ P: 2×2	$3 \times 3 \times 512$ P: 2×2	$2 \times 2 \times 2048$	1024 Dr: 0.1	C

Table 1: **CNN architectures.** Acronyms: 'P'=pooling size; 'Norm'=normalization; 'Dr'=dropout; 'C'=number of outputs (*i.e.*identification $C = 155$, gender $C = 2$, age $C = 1$).

2.2. Main task: people identification

Given a video sequence of a walking person, our main task is to assign a class (*i.e.*identity) to that person from a set of C predefined ones. The last fully-connected layer contains as many units as categories (*i.e.*identities). The loss function for this task is a *softmax log-loss* given by the equation:

$$\mathcal{L}_m(\hat{\mathbf{y}}, c) = -\hat{y}_c + \log \sum_{k=1}^{C} e^{\hat{y}_k}, \quad (2)$$

where $\hat{\mathbf{y}}$ is the output vector of the network for the main task, $\hat{y}_j$ is the j-th component of the output vector and c is the ground-truth class. This loss function favours the generation of high values for the component corresponding to the target class, and low values for the other classes. Therefore, the identity of a given sample $\mathbf{v}$ is derived from $\hat{\mathbf{y}}$ by taking its component whose value is maximum.

2.3. Auxiliary tasks

2.3.1. Gender recognition

Given an input sequence, the goal of this task is to assign it the label either 'female' or 'male' by using gait information, *i.e.*, a binary classification problem where the input is a multidimensional matrix. The last fully-connected layer of the CNN contains two units (one per gender) and the loss function $\mathcal{L}_g$ for this task is a *softmax log-loss* similar to Eq. 2 with $C = 2$.

2.3.2. Age estimation

The goal of this task is to assign it a (positive) real number, *i.e.*the age. The last fully-connected layer of the CNN contains a single unit that will represent the estimated age of the input sample. The loss function for the age task is a regression layer with *'Tukey's biweight loss'* [14]. This is a robust regression loss function that reduces the influence of outliers during the learning process. It has shown good results for the problem of age estimation on faces [15]. Therefore, we adapt it for our task: gait-based age estimation. The age loss $\mathcal{L}_a(\hat{y}_i, y_i) = \rho(r_i^{\text{MAD}})$ is based on the residual r_i of the training samples $r_i = y_i - \hat{y}_i$ and is defined by the equations:

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

$$r_i^{\text{MAD}} = \frac{y_i - \hat{y}_i}{1.4826 \times \text{MAD}} \quad (4)$$

$$\text{MAD} = \underset{k \in \{1,\ldots,S\}}{\text{median}} \left(\left| r_k - \underset{j \in \{1,\ldots,S\}}{\text{median}} (r_j) \right| \right) \quad (5)$$

where constant c is set to 4.6851 as suggested in [15], S is the number of training samples, and the subscripts k and j are indexes referred to the training samples. This layer may be replaced by another regression-like loss function, as the L2 distance: $\mathcal{L}_a(\hat{y}_i, y_i) = ||\hat{y}_i - y_i||_2^2$. However, early experiments showed that Tukey's loss generalized better than L2.

2.3.3. Identity verification

In the identity verification task, given two samples, we aim at deciding whether both samples belong to the same person or not. For this task we design a siamese network (*i.e.*weights are shared between two CNN), where the loss function for the verification task is given by the following equation [16]:

$$V(\mathbf{f}_i,\mathbf{f}_j,y_{ij})=\begin{cases} \frac{1}{2}||\mathbf{f}_i-\mathbf{f}_j||_2^2, & \text{if} y_{ij}=1 \\ \frac{1}{2}\max(0,m-||\mathbf{f}_i-\mathbf{f}_j||_2)^2, & \text{if} y_{ij}=-1 \end{cases}, \quad (6)$$

where $\mathbf{f}_i$ and $\mathbf{f}_j$ are the output of the last fully-connected layer of two gait sequences to be compared. When $y_{ij} = 1$, both sequences belong to the same subject, and the function tries to minimize the L2 distance between the feature vectors. Whereas, when $y_{ij} = -1$, what means different subjects, the loss function prefers a distance between the feature vectors larger than a chosen margin m. During training, m is updated to minimize the verification error on the current batch.

2.4. DMT training

We adopt the *task-wise early stopping criterion* proposed in [17] to detect the auxiliary task that has already converged. For training the networks with back-propagation [18], we use Stochastic Gradient Descent (SGD) with momentum. We use a batch size of 256 samples. The initial learning rate is $\eta = 0.01$, which is kept unchanged during the first 10 epochs and, then, it is reduced by 0.1 when the validation error stops improving. Momentum is set to $\mu = 0.9$ and weight decay to 0.0005. Filters are initialised with random values from a normal distribution with mean 0 and standard deviation 0.01. Bias are all set to zero. The maximum number of epochs is set to 30, although the training may be stopped earlier if the validation error stops improving.

3. EXPERIMENTAL RESULTS

Dataset description. We carry out our experiments on 'TUM Gait from Audio, Image and Depth' [9] (TUM-GAID) dataset, where 305 subjects walk on the following conditions: *normal* walk (*N*), carrying a *backpack* (*B*), wearing coating *shoes* (*S*), as used in clean rooms for hygiene conditions, and *elapsed time* (*TN-TB-TS*). We follow the experimental protocol described in the original paper [9], where 100 subjects are used for training, 50 for validation and 155 for testing. In addition to subject identity, each sequence is labelled with both the gender and the age of the subjects, what will be used to train the auxiliary biometric tasks.

Implementation details. We use the Directed Acyclic Graph (DAG) module of the MatConvNet library [19] for implementing our DMT-CNN models. Apart from the identification task, which requires fine-tuning the classification layer on the test 'subject partition' (test 'sequence partition' [9] are never used for training), the auxiliary tasks do not see any subject from the test 'subject partition' during training. Weights in convolutional layers are frozen during the fine-tuning stage, only fully-connected layers are updated. Given a test sequence, we split it into overlapping subsequences of length 25 frames, with 80% of overlap between subsequences. To increase the number of training samples, we apply 8 spatial displacements of ± 5 pixels in all directions, and include their mirror sequences, obtaining about $270k$ training samples. To obtain results at sequence level for the identification and gender tasks, we use either a) component-wise product of the softmax layers (SMP), *i.e.*product of probabilities; or, b) majority voting on the labels estimated from SVM. For the age estimation task, we use the median of the estimated ages.

Metrics. We use the following metrics in our experiments: *Accuracy (Acc):* in identification, the percentage of samples correctly classified; *Area-Under-the-Curve (AUC):* corresponds to the area under the given curve – when used with the precision-recall curve, values near 1 are preferred – used for the gender and verification tasks; *Mean Absolute Error (MAE):* computes the average on a set of absolute values, it is used on the task age for comparison with [9]; *Accuracy at Equilibrium-Point (EP):* when the percentage of true positives is equal to the percentage of true negatives – values near 100% are preferred – used for gender and verification tasks.

3.1. Results

A comprehensive set of models have been trained for these experiments. Different auxiliary task weights λ_t have been tried in the interval $[0.025, 0.5]$. Due to space limits, only selected results are reported.

Auxiliary tasks speed up convergence of the main task. Fig. 2 shows a comparative of both the energy of the loss function (a,c) and the classification error (b,d) for the identification task. In all the plots, the lower the better. In terms of CNN complexity, we can observe that type *vB* achieves lower loss and training error faster than type *vA*. In our opinion, this means that a larger amount of hidden units enables a better representation of the data distribution. Focusing on CNN type *vA*, Fig. 2.(b) shows that the combination of tasks 'gender' (at 0.1) and 'age' (at 0.1) with 'identification' obtains a training error below 0.1 in epoch 4, whereas the single task of identification 'id' only achieves a value above 0.3. This means that a faster convergence of the objective function is accomplished by adding side-tasks to the main one. However, it seems that these effects are moderated when the complexity of the networks increases (see CNN type *vB*). In terms of AUC, the best choice is adding the 'age' task (with $\lambda_t = 0.1$) to the main one as it achieves the lower AUC in all the cases.

People identification. We compare the identification accuracy of several DMT models in Fig. 3.(a). The baseline case is only 'id'. Note that the best average result is obtained when 'gender' and 'age' tasks are included in the DMT model. In general, adding the 'age' task seems to be a good choice. The scenarios that benefit most from the auxiliary tasks are *B* and *S*, the latter especially with 'verif' task. Moreover, we clearly improve on the average results reported by Hofmann *et al.* [9].

Gender recognition. In this experiment, we train SVM classifiers on top of the L2 normalized signatures (*i.e.*layer 'Full01') obtained from different models. Our baseline case

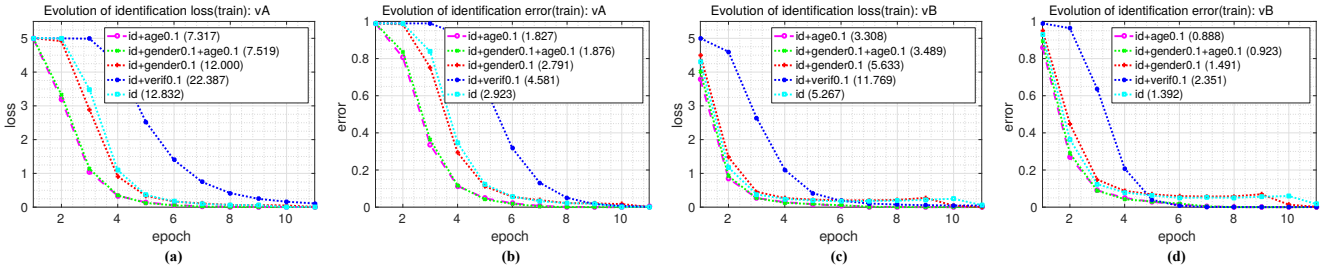

Fig. 2: **Training multi-task networks (v2)**. CNN type *vA*: **(a)** Loss of the identification task. **(b)** Identification error. CNN type *vB*: **(c)** Loss of the identification task. **(d)** Identification error. The area under the curve (AUC) for each case is included in the legend (the lower, the better). Legend format: id+*task&weight*. Best viewed in digital version.

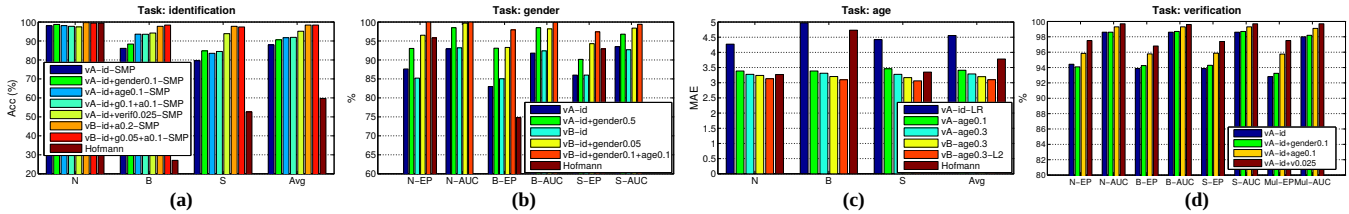

Fig. 3: **Performance on the tasks**. **(a)** Identification. **(b)** Gender. **(c)** Age. **(d)** Verification: bars 'Mul' correspond to the mixed scenario experiment. Notes: equlibrium-point (EP) and AUC ($\times 100$). Legend format: *net+task&weight*. (See main text)

is given by the use of the descriptors directly obtained from the model trained on the single task 'id'. The results on gender recognition, summarized in Fig. 3.(b), suggest that (*i*) adding an explicit loss for gender improves the accuracy of the task, and (*ii*) adding the 'age' task brings further improvements. Moreover, we show that our DMT signatures improve on Hofmann *et al.* [9].

Age estimation. Our baseline is a linear robust regressor (LR) trained on top of the gait signatures (*i.e.*layer 'Full01') obtained from the model with the single task 'id' by using the function *'robustfit'* of MATLAB. In Fig. 3.(c), we can observe that (*i*) adding an explicit loss for regression improves the estimation, and (*ii*) increasing the network complexity and setting a greater weight to the age task reduces the MAE. Note that using L2 loss for regression helps to reduce the error in this task, however, we observed that identification worsened.

Effect of verification loss. We generate 10 random sets of $3k$ pairs of samples each: $1k$ positive (same class) and $2k$ negative (different class). The last fully-connected layer of the DMT-CNN (type *vA*) is used as gait signature (256 dimensions). We report the mean AUC and the mean EP on the 10 sets. This experiment has been carried out both independently per scenario (bars *N* to *S*), and mixing scenarios (bars *Mul*). For the latter case, known in the literature as '*uncooperative gait recognition*' [20], the possible pairs of scenarios are balanced. Our baseline case is the use of the signatures obtained from the single task model for 'id'. The results summarized in Fig. 3.(d) suggest that (*i*) using the verification loss improves the accuracy in this task, even with a small weight ($\lambda_t = 0.025$); and (*ii*) adding either the gender or the age tasks brings improvements on the baseline as well. We investigate as well the contribution of the verification task to the separability of samples by using a linear SVM for identification. As a reference, the accuracy improves from 94.7% with *vA* to 97.8% with *vA*+verif0.25.

Method	N	B	S	*Avg*	TN	TB	TS	*Avg*
Ours-1 (SVM)	100	97.1	97.1	98.1	53.1	59.4	50.0	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**
Hofmann *et al.* [9]	99.4	27.1	52.6	59.7	44.0	6.0	9.0	19.7
RSM [21]	100	79.0	97.0	92.0	58.0	38.0	57.0	51.3
CNN (SVM) [11]	99.7	97.1	97.1	98.0	59.4	50.0	62.5	57.3
PFM [3]	99.7	99.0	99.0	**99.2**	78.1	56.3	46.9	**60.4**

Table 2: **State-of-the-art on Identification.** Recognition accuracy at sequence level. Ours-1: id+age0.2 (*vB*) ; Ours-2: id+verif0.1 (*vB*); 7-NN: 7-Nearest Neighbour with PCA-128.

State-of-the-art comparison. Our results are comparable to the state-of-the-art (Tab. 2), improving on several cases (results taken from the published papers and using the same experimental setup). For our best result ('*Ours-2 (7-NN)*'), the gait signatures are PCA-compressed at 128 dimensions and classified with a 7-Nearest Neighbour classifier. Recall that our CNN is simpler (fewer units and layers) than [11] and our video resolution is eight times lower than the one used in [3].

4. CONCLUSIONS AND FUTURE WORK

This paper has introduced a novel deep multi-task framework for gait-based biometrics using as input low-level motion data (*i.e.*optical flow). Our thorough empirical study on TUM-GAID shows that by combining gait-based tasks in a deep multi-task framework: (*i*) the accuracy of the identification task improves; (*ii*) the convergence speed of the training stage increases; and (*iii*) the CNN filters obtained in the first layers encode spatio-temporal patterns valid for all the involved tasks. This multi-task approach achieves state-of-the-art results on the different tasks by using compact CNN architectures. As future work, we plan to extend this study to other low-level features as RGB and depth, and additional tasks.

5. REFERENCES

[1] Rich Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[2] Ju Han and Bir Bhanu, "Individual recognition using gait energy image," *IEEE PAMI*, vol. 28, no. 2, pp. 316–322, 2006.

[3] Francisco M. Castro, M.J. Marín-Jiménez, N. Guil Mata, and R. Muñoz Salinas, "Fisher motion descriptor for multiview gait recognition," *Intl. J. of Patt. Recogn. in Artificial Intelligence*, vol. 31, no. 1, pp. 1–40, 2017.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, 2016.

[5] C. Yan, B. Zhang, and F. Coenen, "Multi-attributes gait identification by convolutional neural networks," in *Intl. Conf. on Image and Signal Processing (CISP)*, 2015, pp. 642–647.

[6] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep CNNs," *IEEE PAMI*, vol. PP, no. 99, 2016.

[7] T. Wolf, M. Babaee, and G. Rigoll, "Multi-view gait recognition using 3D convolutional neural networks," in *Proceedings of the IEEE International Conference on Image Processing*, 2016, pp. 4165–4169.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[9] Martin Hofmann, Jrgen Geiger, Sebastian Bachmann, Bjrn Schuller, and Gerhard Rigoll, "The TUM Gait from Audio, Image and Depth (GAID) database: Multimodal recognition of subjects and traits," *J. of Visual Com. and Image Repres.*, vol. 25, no. 1, pp. 195 – 206, 2014.

[10] Francisco M. Castro, Manuel J. Marín-Jiménez, and Nicolás Guil, "Multimodal features fusion for gait, gender and shoes recognition," *Machine Vision and Applications*, vol. 27, no. 8, pp. 1213–1228, 2016.

[11] Francisco M. Castro, Manuel J. Marín-Jiménez, Nicolás Guil, and Nicolas Pérez de la Blanca, "Automatic learning of gait signatures for people identification," in *Intl. Work-Conference on Artificial Neural Networks (IWANN)*, 2017.

[12] Gunnar Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. of Scandinavian Conf. on Image Analysis*, 2003, vol. 2749, pp. 363–370.

[13] P. KaewTraKulPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Video-Based Surveillance Systems*, pp. 135–144. 2002.

[14] 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.

[15] Vasileios Belagiannis, Christian Rupprecht, Gustavo Carneiro, and Nassir Navab, "Robust optimization for deep regression," in *ICCV*, 2015, pp. 2830–2838.

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

[17] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Facial landmark detection by deep multi-task learning," in *ECCV*. Springer, 2014, pp. 94–108.

[18] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, no. 3, pp. 1, 1988.

[19] A. Vedaldi and K. Lenc, "MatConvNet – Convolutional Neural Networks for MATLAB," in *Proceedings of the ACM Int. Conf. on Multimedia*, 2015.

[20] Khalid Bashir, Tao Xiang, and Shaogang Gong, "Gait recognition without subject cooperation," *Pattern Recognition Letters*, vol. 31, no. 13, pp. 2052 – 2060, 2010.

[21] Yu Guan and Chang-Tsun Li, "A robust speed-invariant gait recognition system for walker and runner identification," in *Intl. Conf. on Biometrics (ICB)*, 2013, pp. 1–8.