View Resistant Gait Recognition

Anna Sokolova
National Research University
Higher School of Economics
20 Myasnitskaya str., Moscow, Russia;
Samsung-MSU Laboratory, Lomonosov MSU, GSP-1,
Leninskie Gory, Moscow, Russia
+7 (495) 939-01-90
ale4kasokolova@gmail.com

Anton Konushin
Samsung AI Center
5c Lesnaya str., Moscow, Russia;
Samsung-MSU Laboratory, Lomonosov MSU, GSP-1,
Leninskie Gory, Moscow, Russia
+7 (495) 939-01-90
anton.konushin@graphics.cs.msu.ru

ABSTRACT
Human gait is one of the biometric characteristics that a person can be identified by. However, the wide applicability of gait recognition in real life is prevented by a great variety of conditions that affect the gait representation, such as different viewpoints. In this work, we present a novel view resistant approach to overcome the multi-view recognition challenge. The new loss function is proposed to increase the stability of the model to view changes. Besides this, the cross-view embedding of the gait features is made to enhance their discriminant ability which improves the recognition accuracy as well. The proposed approaches show a significant gain in quality and allow to achieve the state-of-the-art accuracy on the most common benchmark and outperform the most successful model on the majority of the views and on average.

CCS Concepts
• Computing methodologies→Object identification; Biometrics

Keywords
Biometrics; gait; neural networks; multi-view recognition.

1. INTRODUCTION
Desire for security is one of the basic human aspirations: people tend to protect themselves, their houses and properties. Nobody wants the thief to get into their car or house or steal something from the bag. The development of video surveillance systems gives the possibility to collect a great amount of video data that can help to find the criminals and prevent committing new crimes. And modern computer vision methods allow to automatize many security problems, such as object detection and recognition. However, the most common recognition characteristic is the face of a person that can easily be hidden or faked (using makeup or mask). Other popular features such as fingerprints or iris require direct interaction with a person which sometimes is impossible. But there is one more biometric index, gait or walking manner, that does not have these disadvantages and, therefore, can be used for contactless automatic recognition.

ICVIP 2019, December 20–23, 2019, Shanghai, China

ACM ISBN 978-1-4503-7682-2/19/12…$15.00
https://doi.org/10.1145/3376067.3376083

Due to physiological studies [6, 13], gait is a unique identifier which cannot be faked. Nevertheless, there is great variability of different conditions that can make gait look differently or change its computer representation. The first type of conditions contains shoes variation, being drunk or sick, or carrying something heavy. The second one includes different lightening, clothes that can change human's figure and hide some body parts, and view variation. The view variation is probably the most complex challenge in gait recognition problem. It is obvious for human's eye that there is the same person captured under different angles, but computer gets two absolutely different video sequences and it is complicated to train the model stable to view changes.

Despite the fact that modern deep learning methods show significant results in most of computer vision problems, gait recognition challenge is still not subdued. Although being a video classification problem, it is close to action recognition, the difference between two gaits is much smaller than between two actions and most of the methods are not transferable from one problem to another.

Most of gait recognition methods tend to use specific features such as silhouette mask [2, 9] or human pose defined by the set of human body key points. Such approaches achieve quite high recognition accuracy, however, to analyze the motion thoroughly one need to know not only the situation in each distinct frame but the information about the dynamics of the body [8]. One of the ways to get such information is considering optical flow between consecutive frames. The approach [17] based on this idea shows really good quality, but similarly to the others, it suffers from lack of view stability.

In this work, we propose a novel view-invariant approach to gait recognition. Being based on one of the state-of-the-art methods [17], it aims to overcome the problem of multiple viewing angles and train the gait features that do not depend on the view. We are the first to use a new loss (View Loss) that can complement conventional loss function and work as a regularizer. Besides this, we propose the cross-view triplet probabilistic embedding (CV TPE) that can be applied as post-processing to get rid of any view dependence. Being applied to [17] each of these two approaches improves the recognition quality, and their union shows that they can complement each other and outperform the state-of-the-art models on the gait recognition benchmark.

2. RELATED WORK
Gait recognition methods have been developing actively recently. Although it is a computer vision problem where the neural networks show excellent results, it is very specific and non-deep methods still compete with the deep ones. The most common

features used for recognition are silhouettes and various aggregations of silhouette sequences [2, 9]. Gait energy images (GEI) [9] are the most popular basic gait descriptors and a great variety of methods are based on them. In [1, 11] different GEI embeddings are proposed based on linear discriminant analysis. The investigations show that the single-view recognition is much easier than the cross-view one, thus, many authors suggest that gait features obtained from different angles should be embedded into one common subspace or transformed into each other. Several works [12, 20] propose to transform the descriptors to a common view. While [12] considers linear transformation, [20] is based on modern popular generative neural network approaches: autoencoders and adversarial networks. These generative models are used to "turn" gait images to the same view and solve a single-view problem. However, the best results until recent time were achieved by neural methods not making any view transformations but computing deep gait representations and considering their similarity. Different Siamese architectures with several streams are investigated in [18, 19, 22] and the ways of stream fusion, feature aggregation and similarity measuring are compared. The first approach able to outperform [19] being state-of-the-art for several years is [5] considering the sets of silhouettes that are invariant to permutation and allow to mix the frames obtained under different views. The authors consider various aggregation functions and their combination (Set Pooling) to get set-level features from frame-level ones and also present Horizontal Pyramid Mapping which splits the neural feature maps into the strips of different scales in order to aggregate multi-scale information.

Although the silhouettes and GEIs are the most popular gait descriptors and allow to achieve high identification results, this data is not the only one that can be used for gait recognition. Another source of information about gait is optical flow (OF). Reflecting the movements of the points between the frames, OF based approaches make the motion of greater interest than appearance. The idea of considering OF is proposed in [15] for action recognition and after its success, it was applied to gait in [4, 16]. In [17] this approach is united with pose-based method: the OF is considered in different parts of the body which allows to pay more attention to some body parts than on the others. This approach achieved state-of-the-art quality and outperformed [19] for several angles. And it is the method we base our model on.

3. PROPOSED METHOD

3.1 Baseline

Let us firstly describe some details of the baseline approach [17].

The pipeline of this method is as follows:

1. Optical flow estimation between consecutive frames;
2. Pose estimation in each frame;
3. Neural features extraction;
4. Feature aggregation and classification.

The first two steps are made independently using the existing methods of optical flow [7] and pose [3] estimation. Having executed both preprocessing steps the patches of OF are cropped from several parts of the body. The authors consider five areas: full body, upper and lower parts of the body and two patches around the feet. Having found the set of body key points the bounding boxes are calculated for these areas and OF patches are cropped to be fed into the network.

The third step requires preliminary neural network training which is one of the most challenging parts of the problem. The network is trained for classification task (by LogLoss minimization), to predict the probability of the input patch to belong to one of the subjects from the training set. WideResNet architecture is chosen since it is not very deep but as residual network shows good training ability. This architecture has 256 neural units on the last hidden layer which is the dimension of future neural representations.

When the network is trained it is used as a feature extractor. The hidden representations are calculated for each patch in each OF map, and then the obtained descriptors are averaged over time and concatenated over the body parts. Such a procedure allows to get one high-dimensional descriptor for one video that can further be normalized and classified by Nearest Neighbor method. PCA decomposition can be applied to the descriptors prior to aggregation to get rid of extra noise and accelerate the classification.

In this work, we propose two important modifications of this pipeline. The first one concerns the training process: having the same architecture we add an auxiliary loss function called View Loss working as a regularizer to increase view stability. The second modification is made on the last step: instead of direct feature aggregation and classification, we train a cross-view embedding which improves the discriminant ability of the descriptors. The scheme of the baseline algorithm and the modifications are shown in Fig. 1 and the details are discussed in the following sections.

3.2 View Loss

Aiming to recognize the subject independent of viewing angle, we need the trained features of the same subject to be similar for different views. Thus, we propose to consider special view-resistant loss function while neural network training.

The network used for feature extraction is trained for classification task to predict the probability distribution over the set of training subjects. However, additionally to classical LogLoss, we propose an auxiliary loss called "view loss" to regularize the model and decrease the overfitting. It aims to make the hidden representations of videos of the same subject close to each other even if they have different viewing angles. Thus, it penalizes for the big difference between the representations of videos with one certain view and averaged representation of videos with different views.

In details, let i be an Id of a subject, α be an angle. Let us consider two data batches: the first one $\{b_{i,\alpha}\}$ consists of the data for subject i with view α, and the second one $\{b_i\}$ consists of data for the same subject i but captured under any angle (angles in the batch can be different). Let $d_{i,\alpha}$ and d_i be the average of last hidden layer outputs of batches $b_{i,\alpha}$ and b_i, respectively.

Since we use the hidden representations as the gait features and find the nearest one in the database, we aim to make the features of the same subject closer to each other regardless of the view. Hence, they should be close to the mean value over different views, and we need to bring $d_{i,\alpha}$ and d_i closer. To achieve it, we consider such pairs of batches and the following loss function for them:

$$L_{reg} = \lambda L_{view} + L_{clf},$$

where $L_{view} = \left\| d_{i,\alpha} - d_i \right\|^2$ is the described view loss and $L_{clf} = LogLoss(b_{i,\alpha}) + LogLoss(b_i)$ is cross-entropy for classification applied to both batches. Thus, we add the view term to classical loss function which can be considered as the regularization to

prevent view memorizing. Since we need to sample such batches in a special way (the subject Id is fixed inside the batch and the angles are sometimes fixed, as well), we cannot train the network only on such batches. So, we alternate the conventional optimization steps makes "view optimization" once in k steps. The whole algorithm for one epoch appears to be as follows:

for j in range (epoch size) **do**
  Sample random batch;
  Make optimization step: $L_{clf} \to min$
  if j mod $k = 0$ **then**
  Sample random subject i, random view α;
  Sample batches $\{b_{i,\alpha}\}$, $\{b_i\}$;
  Make regularized optimization step: $L_{reg} \to min$
  end if
end for

Such regularized optimization steps obviously increase the unregularized loss but prevent the fall into the local minimum. Being applied to the network that has already stopped being optimized, the described optimization process can help to leave current minimum if it is not optimal and then descent to "real" global minimum.

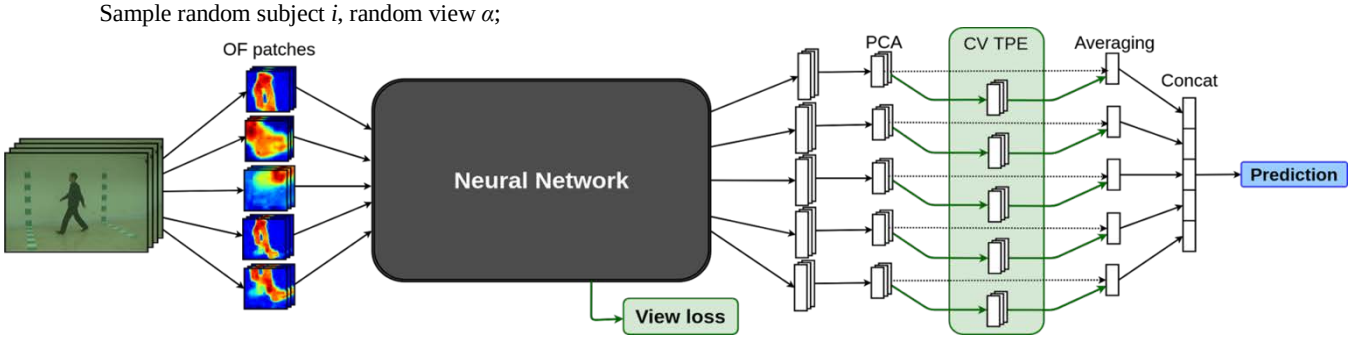

Figure 1. The pipeline of the algorithm. The green boxes (view loss and CV TPE) correspond to our novelties.

3.3 Cross-View Triplet Probabilistic Embedding

Although the training process described in Section 3.2 aims to get rid of view dependence, we propose one more approach that can be applied to the neural features to increase view resistance more. This approach is based on Triplet Probabilistic Embeddings (TPE) [14] that was proposed for verification improvement. Despite the fact that our main goal is classification, we, however, apply this method to the neural features obtained from the network.

TPE aims to find an embedding (projection matrix) to make features of the same object closer to each other than the features of different objects. It is trained to find the embedding such that

$$S_W(v_a, v_p) > S_W(v_a, v_n) \quad (1)$$

where $S = S_W$ is similarity which is usually defined as cosine measure, W is a parametrization of the embedding, features v_a, v_p belong to the same object, and v_n to the other one.

The inequality (1) is quite abstract, and the following optimization problem is formulated to achieve it. Let us consider the probability of the triplet (v_a v_p, v_n):

$$p_{apn} = \frac{e^{S_W(v_a,v_p)}}{e^{S_W(v_a,v_p)} + e^{S_W(v_a,v_n)}}$$

The closer v_a and v_p are relative to the similarity of v_a and v_n the higher this probability p_{apn} is. Thus, in order to get the optimal parameters W, we can follow the maximum likelihood method and maximize this probability or its logarithm which is usually easier for optimization.

$$Loss = \sum_{(v_a,v_p,v_n)} -log\,(p_{apn}) \to \min_W.$$

However, TPE was proposed without respect to multi-view recognition and does not take angles under consideration. We propose its modification to use the information about the view while training and make the features of the same subject with different angles closer to each other.

3.4 Cross-view modification

We want the features of the same object captured under different views to be close to each other, ideally, closer than the features of different objects captured under the same angle. Thus, we get the following condition to be satisfied:

$$S_W(v_{a,\alpha}, v_{a,\beta}) > S_W(v_{a,\alpha}, v_{n,\alpha}),$$

where $v_{a,\alpha}$, $v_{a,\beta}$ are the features of the same object a captured under different views ($\alpha \neq \beta$), and $v_{n,\alpha}$ corresponds to the other object captured under the same view α as $v_{a,\alpha}$.

One more difference from the initial method is that we use Euclidean distance instead of cosine similarity. We get

$$S_W(u, v) = -\|Wu - Wv\|^2,$$

and likelihood maximization leads to optimization problem

$$\sum_{(v_{a,\alpha},v_{a,\beta},v_{n,\alpha})} log \frac{e^{-\|Wv_{a,\alpha} - Wv_{a,\beta}\|^2}}{e^{-\|Wv_{a,\alpha} - Wv_{a,\beta}\|^2} + e^{-\|Wv_{a,\alpha} - Wv_{n,\alpha}\|^2}} \to \max_W,$$

that can be solved by stochastic gradient descent method.

Constructing such an embedding, we make the representations less view-dependent and more id-dependent. The features of the same subject get closer even being obtained under different views.

We additionally implement hard negative mining and for each anchor object choose the closest feature among the set of negatives with the same view.

4. EVALUATION

In this section, we describe the conducted experiments and present the evaluation of the proposed method.

4.1 Dataset

We use CASIA Gait Dataset B [21] as the benchmark since it contains large variability of views and it is the only multi-view

gait dataset which is fully distributed in form of RGB videos, thus, any computer vision method can be applied to this data. The other popular multi-view gait databases, OU-ISIR collections [10], are available only in a silhouette form which prevents applying optical flow based methods to this data and, hence, evaluating the proposed approach. CASIA dataset contains data for 124 subjects, but the presence of 10 videos for each subject for each of 11 different views makes this database large enough.

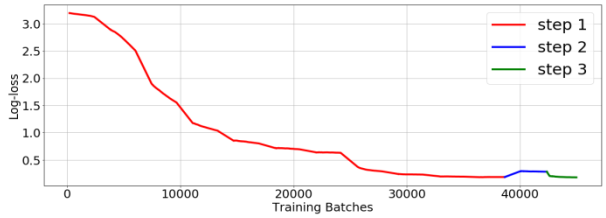

Figure 2. The curve of the training loss during three steps of training process.

Each video in this collection lasts 3-4 seconds and contains a walk of one person without any occlusions. Despite these "simple" conditions the small resolution (320 x 240) and view variability make CASIA database really challenging, and the state-of-the-art quality on this dataset is still far from perfect.

4.2 Experiments and Results

To evaluate the influence of both proposed approaches we have conducted separate experiments for them and then united the models to check if View Loss and CV TPE are interchangeable.

To compare our results with the baseline, we use one of the existing protocols proposed for CASIA database: the network is trained on the first 24 subjects and the rest 100 are used for testing. While testing we fit the classifier on the first 4 videos for each subject and test on the others. Similarly to other classification problems, we measure the accuracy of algorithms, the results of all the experiments are presented in Table 1.

4.2.1 View loss

As mentioned in Section 3.2, view loss acts as a regularization to prevent getting into a local minimum. However, firstly the network has to be trained and achieve any "optimal point". Thus, we have implemented the following training process:

1. Unregularized training (until LogLoss stops decreasing);
2. Regularized training with $\lambda = 10^{-3}$, $k = 20$ (until both classification and view losses stop changing);
3. Unregularized training (until LogLoss stops decreasing).

Despite the fact, that training loss has increased after the step 2 the accuracy on testing part after this step is higher for many angles. During step 3 the loss on training set has decreased and achieved the smallest value, and the final accuracy on testing data is the highest after this step. The evaluation is made for each pair of gallery and probe angles, but for simplicity and convenience, we present the averaged results for cross-view recognition for four probe angles 0°, 54°, 90° and 126° (the average is calculated over ten angles different from the probe one).

As we supposed, the model got into the local minimum after the first step of training, and the regularized training procedure has "pulled" the model from this local minimum and moved in the direction of global minimum in order to get lower loss after the last descent (Fig. 2). It is hardly noticeable on the graph, but actually, the loss after the third step is 0.18 which is 13% lower than 0.21 after the first step. The decrease of the training loss offers the hope that the quality on test set improves as well.

4.2.2 CV TPE

Cross-view triplet probabilistic embedding is trained separately from the network, based on any set of feature vectors. We have trained the embedding on two models: the baseline one [17] and view loss based, to check if both approaches worth applying or they are interchangeable. We have also trained conventional TPE to evaluate the influence of "cross-view" component. In all the cases the embedding is trained on neural features for 24 training subjects (the same as for network training) and then applied to the test set.

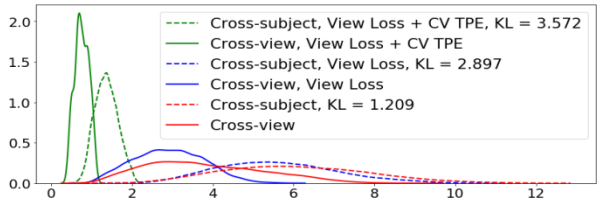

Figure 3. Probability density curves of cross-view (solid) and cross-sub (dashed) distances of baseline model (red), model with view loss (blue) and with both view loss and CV TPE (green).

Prior to measuring the recognition accuracy, we have verified if the features of the same subject really get closer to each other than the features of different subjects with the same view. We have calculated the cross-subject and cross-view distances between different video descriptors (having fixed the view or the subject, respectively), and estimated the probability densities in each case. To estimate the similarity of the distributions we calculate Kullback–Leibler (KL) divergence between the distributions. It measures difference between two distributions: the closer distributions are, the smaller KL value is. The density curves and the values of KL divergence are presented on Fig. 3. As expected, the green pair of distributions (corresponding to both approaches application) has the greatest KL value, which means that these two distributions are the most distant from each other. The red curves correspond to distributions obtained from initial model, they are closer to each other and have the least KL-divergence.

Table 1 shows the comparison of all the models with state-of-the-art approaches and demonstrates the superiority of the combination of proposed approaches over their separate usage. One can see, that View Loss increases the average accuracy by 5.4 percentage points (from 71.6% to 77%) and additional cross view embedding increases the quality by another 2.4 percentage points. The result turns out to be 1.2 percentage points higher than the average accuracy of state-of-the-art approach [5]. It confirms that two proposed concepts are not interchangeable and complement each other allowing to outperform other existing methods.

4.3 Implementation Details

The pipeline was implemented using several public Python libraries: OpenCV methods were chosen for OF estimation and Open Pose approach [3] for pose estimation. The neural networks were implemented in PyTorch framework, which allowed us to improve the initial (step 1) quality comparing to the baseline model. The model has been training 4 hours on NVIDIA GTX 1070 GPU.

5. CONCLUSIONS

In this study, we addressed the problem of multi-view gait recognition. We have presented two novel approaches that increase gait recognition view stability. Actually, the proposed view loss can be integrated into any neural architecture, and the

embedding is applicable to any model prior to feature classification, thus, both concepts are model-free and can be applied to any neural network based model.

The experiments show that both approaches improve the cross-view recognition accuracy of the baseline algorithm and their combination applied to the body part based model allows to achieve the state-of-the-art quality.

Table 1. Comparison of cross-view recognition accuracy of our approach and state-of-the-art models

	Average accuracy [%]				
Method	0°	54°	90°	126°	Avg
View loss, step 1	59.7	80.1	68.9	77.7	71.6
View loss, step 2	53.4	79.8	70.3	81.4	71.2
View loss, step 3	64.4	81.8	72.6	81.1	75.0
View loss, step 3 + normalization	65.9	83.6	74.6	83.7	77.0
Part-based [17] + TPE	56.9	82.6	73.5	82.4	73.9
Part-based [17] + CV TPE	62.6	84.6	75.6	84.4	76.8
View loss, step 3 + CV TPE + normalization	**69.3**	86.3	**75.8**	**86.0**	**79.4**
Part-based [17]	-	77.8	68.8	74.7	-
Wu [19]	54.8	77.8	64.9	76.1	68.4
GaitSet [5]	64.6	**86.5**	75.5	**86.0**	78.2

6. REFERENCES

[1] K. Bashir, T. Xiang, and S. Gong. 2010. Gait recognition without subject cooperation. *Pattern Recognition Letters*, 31, (Oct. 2010), 2052–2060. DOI=10.1016/j.patrec.2010.05.027

[2] K. Bashir, T. Xiang, and S. Gong. 2009. Gait recognition using gait entropy image. In *Proc. of 3rd international conference on crime detection and prevention*, IET, London, UK, 1–6. DOI= 10.1049/ic.2009.0230

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. 2017. Real time multi-person 2d pose estimation using part affinity fields. In *Proc. CVPR*, IEEE, NY, 1302–1310. DOI=10.1109/CVPR.2017.143

[4] F. M. Castro, M. J. Marín-Jiménez, N. Guil, and N. Pérez de la Blanca. 2016. Automatic learning of gait signatures for people identification. *Advances in Computational Intelligence*, (Mar. 2016), 257–270. DOI=10.1007/978-3-319-59147-6_23

[5] H. Chao, Y. He, J. Zhang, and J. Feng. 2019. GaitSet: Regarding gait as a set for cross-view gait recognition. In *AAAI*. AAAI Press, Palo Alto, CA, 8126–8133. DOI=10.1609/aaai.v33i01.33018126

[6] J. E. Cutting and L. T. Kozlowski. 1977. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, 9, 5, (May 1977), 353–356. DOI= 10.3758/BF03337021

[7] G. Farnebäck. 2003. Two-frame motion estimation based on polynomial expansion. In *Image Analysis*, (May 2003), 363–370. DOI=10.1007/3-540-45103-X_50

[8] Y. Feng, Y. Li, and J. Luo. 2016. Learning effective gait features using LSTM. In *INT C PATT RECOG*, IEEE, New York, NY, 325–330. DOI=10.1109/ICPR.2016.7899654

[9] J. Han and B. Bhanu. 2006. Individual recognition using gait energy image. IEEE *TPAMI*, 28, 2, (Mar. 2006), 316–322. DOI= 10.1109/TPAMI.2006.38

[10] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi. 2012. The OU-ISIR Gait Database Comprising the Large Population Dataset and Performance Evaluation of Gait Recognition. *IEEE Trans. on Information Forensics and Security*, 7, 5, (Oct. 2012), 1511–1521. DOI=10.1109/tifs.2012.2204253

[11] A. Mansur, Y. Makihara, D. Muramatsu, and Y. Yagi. 2014. Cross-view gait recognition using view-dependent discriminative analysis. In *Proc. of International Joint Conference on Biometrics*, IEEE, New York, NY, 1–8. DOI=10.1109/btas.2014.6996272

[12] D. Muramatsu, Y. Makihara, and Y. Yagi. 2015. View transformation model incorporating quality measures for cross-view gait recognition. *IEEE Transactions on cybernetics*, 46, (Jul. 2015), 1602–1615. DOI=10.1109/tcyb.2015.2452577

[13] M. P. Murray. 1967. Gait as a total pattern of movement. *American Journal of Physical Medicine & Rehabilitation*, 46, 1, (Feb. 1967), 290–333.

[14] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa. 2016. Triplet probabilistic embedding for face verification and clustering. In *Proc. of the 8th International Conference on Biometrics Theory, Applications and Systems*, IEEE, New York, NY, 1–8. DOI=10.1109/BTAS.2016.7791205

[15] K. Simonyan and A. Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Proc. of the 27th NIPS*, MIT Press, Cambridge, MA, 568–576.

[16] A. Sokolova and A. Konushin. 2017. Gait recognition based on convolutional neural networks. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W4, (May 2017), 207–212. DOI=10.5194/isprs-archives-XLII-2-W4-207-2017

[17] A. Sokolova and A. Konushin. 2019. Pose-based deep gait recognition. *IET Biometrics*. 8, (Mar. 2019), 134–143. DOI=10.1049/iet-bmt.2018.5046

[18] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. 2017. On input/output architectures for convolutional neural network-based cross-view gait recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29, (Oct. 2017), 2708–2719. DOI=10.1109/TCSVT.2017.2760835

[19] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan. 2017. A comprehensive study on cross-view gait based human

identification with deep cnns. *IEEE TPAMI*, 39, 2, (Feb. 2017), 209–226. DOI= 10.1109/TPAMI.2016.2545669

[20] S. Yu, H. Chen, E. B. G. Reyes, and N. Poh. 2017. Gaitgan: Invariant gait feature extraction using generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition Workshops(CVPRW)*, IEEE, New York, NY, 532–539. DOI=10.1109/CVPRW.2017.80

[21] S. Yu, D. Tan, and T. Tan. 2006. A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition. In *INT C PATT RECOG*. IEEE, New York, NY, 441–444. DOI=10.1109/icpr.2006.67

[22] C. Zhang, W. Liu, H. Ma, and H. Fu. 2016. Siamese neural network based gait recognition for human identification. In *Proc. Of International Conference on Acoustics, Speech and Signal Processing*, IEEE, New York, NY, 2832–2836. DOI= 10.1109/ICASSP.2016.7472194