MULTI-VIEW GAIT RECOGNITION USING 3D CONVOLUTIONAL NEURAL NETWORKS

Thomas Wolf, Mohammadreza Babaee, Gerhard Rigoll

Technische Universität München
Institute for Human-Machine Communication
Theresienstrae 90, 80333 München, Germany

ABSTRACT

In this work we present a deep convolutional neural network using 3D convolutions for Gait Recognition in multiple views capturing spatio-temporal features. A special input format, consisting of the gray-scale image and optical flow enhance color invaranice. The approach is evaluated on three different datasets, including variances in clothing, walking speeds and the view angle. In contrast to most state-of-the-art Gait Recognition systems the used neural network is able to generalize gait features across multiple large view angle changes. The results show a comparable to better performance in comparison with previous approaches, especially for large view differences.

Index Terms— Deep Learning, Convolutional Neural Networks, Gait Recognition

1. INTRODUCTION

Automated person recognition based on visual cues is a large research area in computer vision. Important applications are surveillance systems in public spaces to increase safety. The most popular approaches use face, iris or fingerprint information for detection and recognition. These methods work well in many applications, but are sometimes impractical. They are sensitive to occlusion, large distances or low resolution data and often require cooperation of the subject.

Gait Recognition identifies people depending on their natural walking motion. A humans' silhouette and gait are coarse features and therefore robust to noise and low resolution. In [1] it was shown that humans can distinguish humanoid locomotion from other motion patterns. Nonetheless, Gait Recognition is a challenging task. The natural walking style and the appearance of a person can be influenced by many factors, such as wearing different clothing or shoes, having an injury or carrying an object. Additional challenges arise with different walking speeds, viewing angles or environmental influences (e.g. walking in stormy or snowy weather).

In this work we present an approach to tackle these challenges in Gait Recognition adapting recently developed concepts in deep learning. A 3D Convolutional Neural Network (CNN) is presented using spatio-temporal information, trying to find a general descriptor for human gait invariant for view angles, color and different walking conditions.

In the following section, related work in the field of Gait Recognition and deep learning is reviewed. Section 3 presents the methods we developed for this approach. In Section 4, the experiments are explained, followed by the results. Finally, in Section 5, we draw our conclusion.

2. RELATED WORK

Gait Recognition approaches can generally be separated into the two categories: 1) model-based and 2) appearance-based techniques. In the former, parameters for a pre-defined model are adapted and in the latter handcrafted gait features are extracted from images or videos. Especially for low resolution videos finding and optimizing an accurate 3D model for view invariance is hard and error-prone. Therefore, we focus on appearance-based models. Liu *et al.* [2] use frieze patterns combined with dynamic time warping for gait sequence matching with similar viewpoints. Kale *et al.* [3] apply Hidden Markov Models for classification on a persons silhouette. In Han *et al.* [4], the averaged silhouette over a complete gait cycle was used as a simple but effective feature, referred to as Gait Energy Image (GEI). Hofmann and Rigoll [5] enhanced the GEI further using gradient histograms, body part localization and α-matte segmentation resulting in the so-called α-pb-GHEI. The approaches [6], [7] and [8] apply a view transformation model transforming gait sequences into desired view points for comparison. In [9], a common subspace for different view angles or types of clothing is learned. Visual-hull based methods create a 3D gait volume from images sequences of multiple temporally synchronized cameras. This model can then be projected into the desired view plane for classification. However, the need for multiple synchronized cameras is often impractical in many real-world scenarios.

In image classification, deep CNN architectures [10], [11] set new benchmarks on popular datasets such as the ImageNet competition. Recent approaches advanced from 2D image classification to 3D video classification. Karpathy *et al.* [12] use a multi-resolution, fovea architecture applying 3D convolutions on different time frames of a video. In the work of Simonyan and Zisserman [13], two CNNs, one operating on individual RGB images and one on optical flow, were designed and obtained good results by fusing together their softmax activations using an SVM. Donahue *et al.* [14] developed a hybrid architecture, concatenating a CNN with a Long Short-Term Memory (LSTM) network, where the CNN embeds the single frames into feature vectors and the LSTM classifies sequences of these vectors. Similar to [12], Tran *et al.* [15] designed a CNN using 3D convolutions, but with a deeper structure fully exploiting spatio-temporal features for video classification. The next section describes how we combine the ides presented in [11], [13] and [15] to come up with a 3D Convolutional Neural Network for Gait Recognition.

3. NETWORK ARCHITECTURE

The concept behind image classification with CNNs can be transferred to video classification by extending the convolutional operation to the temporal domain. An attempt is already presented in [12], but [15] showed the full potential of using 3D convolutions throughout a network for activity recognition. The findings in action recognition for video data serve as inspiration for the strategies in the approach presented in this section.

3.1. Convolutional Filter

As proposed in [15], 3x3x3 convolutional filters with zero padding are used in all convolutional layers in the network (where HxWxT, stands for spatial height, spatial width, and temporal extent). This allows detection of movements in all directions and includes future and past temporal information, using the smallest possible filter size. As illustrated in Fig. 1 stacking multiple convolutional layers with a size of 3x3xT covers the same spatial region as one larger filter (e.g. 5x5xT or 7x7xT, see AlexNet [10]) with a reduced number of parameters. Besides reducing the computational load, it also allows the introduction of more non-linearities in between the additional convolutional layers.

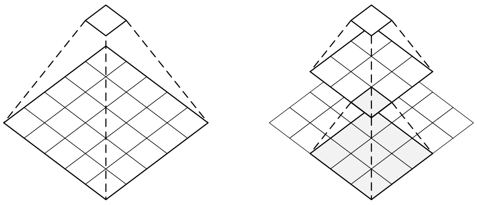

Fig. 1. Comparison between using one 5x5 filter (left, 25 parameters) and two stacked 3x3 filter (right, 18 parameters).

3.2. Nonlinearity and weight initialization

The non-linearities are implemented as Rectified Linear Units (ReLUs) according to Nair and Hinton [16]. Using the activation function $f(x) = \max(0, x)$, ReLUs have multiple benefits. 1) Tackling the vanishing gradient problem for positive inputs, due to a derivative of 1. 2) Being zero for negative inputs can be seen as a regularizer, similar to using dropout. 3) ReLUs need no exponential computation. In order to further improve the learning time, the weights of the convolutional layers are initialized as proposed in [17].

3.3. Topology and preprocessing

Following the design principles of [11], the network architecture is presented in Fig. 2. We set the frame length to 16 as a trade-off between capturing enough temporal information and computational complexity. This temporal depth also showed to work best for activity recognition in [15]. The pooling layers generally have the dimension 2x2x2, with an exception of layer 1 and 3, which have a temporal extent of 1 to avoid collapsing the temporal information too early. As mentioned before, all convolutional layers use 3x3x3xN dimensional convolutional filters, where N is the number of channels (3, 64, 128, 128, 256, 256 and 512 from layer 1 to 7). The features produced by the last convolutional layer are the input for two consecutive fully connected layers with 4096 neurons each implementing dropout with a value of 0.5. The final layer applies the softmax function producing a probability distribution over all classes for classification.

Color and clothing invariance are important aspects in any Gait Recognition algorithm, the goal being to recognize people depending on their gait and not on their clothes. A strength of deep neural networks is the ability to utilize large amounts of data including lots of variance to find a generalized model. Unfortunately, the available datasets include only one to two clothing conditions limiting the ability to learn color invariance. An ideal dataset would include multiple sequences for the same subject with different clothing, enabling the network to learn gait features independent of color or clothing. Tackling this problem a special input format is designed. The first channel of the input *image* is the RGB-image converted to grey-scale. For the second and third channel the optical flow in x and y directions are computed according to [18]. This idea is inspired by the two-stream convolutional network proposed by Simonyan and Zisserman [13], who showed a positive impact of optical flow for video classification. We use optical flow to enhance the capability of the network to learn gait features instead of associating color/clothing a the subject.

In contrast to [15], where the videos are split into non-overlapping sequences, in our approach we use sequences overlapping within test or training set. By cutting the original sequences into non-overlapping training clips of 16 frames one subject may accidentally occur disproportionately often with a similar pose in the first frame and the network learns to associate this starting pose with the subject. In order to avoid this behavior, a video with 50 frames would be split into the clips (1-16), (2-17),... up to frames (35-50).

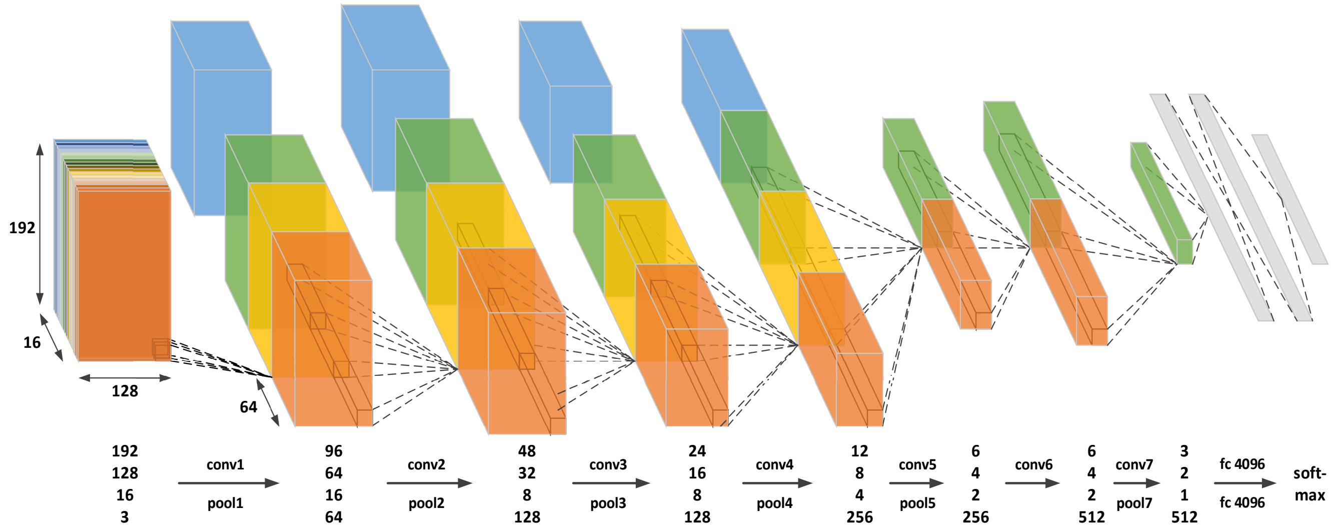

Fig. 2. Topology of the network. All pooling layers are max-pooling layers with a size of 2x2x2, except for pool1 and pool3 which are of size 2x2x1. After each pooling layer a ReLU-nonlinearity [16] is implemented. The features aquired in conv7 are used as inputs of two consecutive fully connected layers with 4096 units each applying dropout with a value of 0.5. The final softmax layer produces a probability distribution over all classes.

4. EXPERIMENTS

4.1. Datasets

The approach is evaluated on three different datasets, namely the CMU Motion of Body (MoBo) [19], the USF Gait Based Human ID Challenge [20] and Casia-B [21]. These datasets contain different walking speeds, clothing conditions and different view angles to evaluate the invariance for different conditions of the features found by the network. The CMU dataset consists of 25 different individuals walking on a treadmill including three different conditions (slow walk (S), fast walk (F) and carrying a ball (B)). In the USF dataset there are 122 subjects walking outside on an elliptical path. The subjects wear two types of shoes, walk on two different surfaces and can either carry a briefcase or not. Additionally the clips were filmed on two separate days (May and November) which caused the people to wear different clothing. However, not all subjects were filmed under all conditions. The Casia-B database contains sequences of 124 subjects walking along a straight line in an indoor environment, recorded from 11 different angles (0°, 18°, ... 180°).

4.2. Test setup

In many experiments training and tests sets have been recorded under different conditions (e.g. *slow walk* for training and *fast walk* for testing), which does not suit a deep learning approach. For conventional methods features can be particularly designed to cope with this setup. However, using CNNs, the features are not designed by hand, but learned from data. The network cannot learn features that are invariant to walking speed if the training data only provides samples of one walking speed. Under this premise, the training/test splits are altered for this approach. In case where training and test data have been recorded under different conditions, both are split separately into a 66% and 33% partition. Then the 66% of the training data are combined with the 66% of the test data to form the altered training data. The remaining 33% of the original training and test data are combined to form the altered test data. Using this type of split, the network is able to learn a representation independent of the view, speed and clothing conditions. Additionally, it enhances the amount of training data improving the performance of the network.

For the CMU dataset, 9 experiments were conducted, 3 using the same conditions for training and testing, and six with different conditions using the altered splits. The USF dataset consists of 12 experiments, where one training set and 12 test sets are defined. The exact conditions can be found in [20]. All 12 experiments use the altered training/test splits in our approach. In the Casia-B dataset each subject performed six walking sequences for all 11 view angles. The experiments are designed to test view invariance, using the six sequences recorded from a 90° view angle for training. For each of the other view angles one experiment is conducted, using the six sequences as test data. The experiment using 90° as test data, uses only four sequences for training and two for testing.

The network is trained with stochastic gradient descent using an initial learning rate of 10^{-4} (Casia-B, USF) to 10^{-5} (CMU) with a momentum of 0.9 and a weight decay of $5 * 10^{-4}$. The learning rate is steadily decreased by a factor of 10, once the loss stalls for one epoch.

4.3. Results and Discussion

The results are presented in Tables 1-3. The experiment, conducted on the CMU Mobo dataset can be seen in Table 1. The first column shows the data split, where the first letter indicates the training set and the second one the test set. It can be seen that for all cases the network is able to find a joint representation independent of walking speed or carrying an object. Due to the special training/test splits, in the last six rows, the results for this approach are duplicated from only three experiments (because S/F = F/S etc.). Despite, the different training conditions, which make a direct comparison difficult, the network shows a very good performance, especially for the experiments with varying conditions.

Exp.	Ve [22]	Liu [2]	Sun [23]	Our
S/S	**100**	**100**	**100**	99
F/F	**100**	**100**	**100**	99
B/B	92	**100**	**100**	**100**
S/F	80	**100**	84	99
F/S	84	84	88	**99**
S/B	48	81	78	**100**
B/S	68	50	80	**100**
F/B	48	50	68	**100**
B/F	48	50	72	**100**

Table 1. Accuracy for the experiments on the CMU Mobo Database in percent. The first column shows the training/test partition.

In Table 2 the results for the USF database can be seen. While the performance for several experiments is slightly below state-of-the art, in four experiments the network shows a significantly better performance. Especially for the last two experiments including different clothing conditions, the results indicate that the network learned a representation invariant for clothing. The otherwise slight decrease in performance compared to previous techniques can be attributed to noise introduced by the outside recording and probably a lower resolution compared to the other two datasets. However, both of these problems can be tackled by increasing the database size, enabling the network to learn the unimportance of the background.

Table 3 shows the performance on the Casia-B dataset, where first column indicates the test view angle. In the altered split training and test data contain both, 90° data and data from the test view angle. Again the network produces a high accuracy across all view angles, indicating a general representation for human gait is found, which can be used for classification.

Overall the performance of the network is very good for all tested datasets underlining the potential of a deep learning approach in Gait Recognition.

Exp.	Sa [20]	Ka [3]	Ha [4]	Hf [5]	Our
A	73	89	90	**99**	89
B	78	88	91	**94**	84
C	48	68	**93**	91	90
D	32	35	90	**93**	83
E	22	28	64	**90**	78
F	17	15	25	64	**81**
G	17	21	36	45	**83**
H	61	85	64	**99**	83
I	57	80	70	**98**	86
J	36	58	70	**96**	78
K	3	17	6	18	**76**
L	3	15	15	21	**80**

Table 2. Test results for the Human ID Gait Challenge. The table shows the accuracy in percent. The conditions for the probe settings can be found in [20].

Exp.	Sa	Liu	Zhe	Ku	Ku	Our
	[21]	[9]	[6]	[7]	[24]	
0°	0.4	20.5	-	-	-	**96.3**
18°	2.4	35.5	-	-	-	**98.2**
36°	4.8	56.5	-	-	-	**98.5**
54°	17.7	81.5	31	-	-	**95.4**
72°	82.3	96.5	60	97	86.3	**94.3**
90°	97.6	-	89	-	95.4	**99.9**
108°	82.3	96.0	89	96	83.3	**98.6**
126°	15.3	89.5	60	-	-	**97.0**
144°	5.2	50.0	-	-	-	**97.4**
162°	3.6	34.5	-	-	-	**99.2**
180°	1.2	21.5	-	-	-	**96.1**

Table 3. Accuracy for the experiments conducted on the Casia-B Dataset in percent. The dashes indicate no available information for certain test cases.

5. CONCLUSION

A new approach has been presented to tackle the challenges in the field of Gait Recognition. View, clothing and walking speed invariance make Gait Recognition a versatile and difficult task. A modern state-of-the art technique using Convolutional Neural Networks is proposed, extracting spatio-temporal features for classification. This representation results in a high accuracy across experiments on different popular databases pointing out the high potential of CNNs for Gait Recognition. Nevertheless, due to the small amount variance and the small database size overall, overfitting is a potential problem. Besides better hardware and bigger network structures for increased performance, a possible solution can be seen in the growing amount of data and the larger databases to come. Using databases including thousands of subject with a large variance in walking behavior and appearance can further boost performance and reduce overfitting.

6. REFERENCES

[1] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Attention, Perception, & Psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[2] Y. Liu, R. Collins, and Y. Tsin, *Gait sequence analysis using frieze patterns*, Springer, 2002.

[3] A. Kale et al., "Identification of humans using gait," *Image Processing, IEEE Transactions on*, vol. 13, no. 9, pp. 1163–1173, 2004.

[4] J. Han and B. Bhanu, "Individual recognition using gait energy image," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 2, pp. 316–322, 2006.

[5] M. Hofmann and G. Rigoll, "Exploiting gradient histograms for gait-based person identification," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE, 2013, pp. 4171–4175.

[6] S. Zheng et al., "Robust view transformation model for gait recognition," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 2073–2076.

[7] W. Kusakunniran, Q. Wu, J Zhang, and H. Li, "Gait recognition under various viewing angles based on correlated motion regression," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 6, pp. 966–980, 2012.

[8] D. Muramatsu, A. Shiraishi, Y. Makihara, M. Z. Uddin, and Y. Yagi, "Gait-based person recognition using arbitrary view transformation model," *Image Processing, IEEE Transactions on*, vol. 24, no. 1, pp. 140–154, 2015.

[9] L. Nini, L. Jiwen, and T. Yap-Peng, "Joint subspace learning for view-invariant gait recognition," *Signal Processing Letters, IEEE*, vol. 18, no. 7, pp. 431–434, July 2011.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] A. Karpathy et al., "Large-scale video classification with convolutional neural networks," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1725–1732.

[13] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, 2014, pp. 568–576.

[14] J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," *arXiv preprint arXiv:1411.4389*, 2014.

[15] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "C3d: generic features for video analysis," *arXiv preprint arXiv:1412.0767*, 2014.

[16] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[17] K. He et al., "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *arXiv preprint arXiv:1502.01852*, 2015.

[18] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2432–2439.

[19] R. Gross and J. Shi, "The cmu motion of body (mobo) database," 2001.

[20] S. Sarkar et al., "The humanid gait challenge problem: Data sets, performance, and analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 2, pp. 162–177, 2005.

[21] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. IEEE, 2006, vol. 4, pp. 441–444.

[22] A. Veeraraghavan, A. K. Roy-Chowdhury, and R. Chellappa, "Matching shape sequences in video with applications in human movement analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 12, pp. 1896–1909, 2005.

[23] B. Sun, J. Yan, and Y. Liu, "Human gait recognition by integrating motion feature and shape feature," in *Multimedia Technology (ICMT), 2010 International Conference on*. IEEE, 2010, pp. 1–4.

[24] W. Kusakunniran, "Recognizing gaits on spatio-temporal feature domain," *Information Forensics and Security, IEEE Transactions on*, vol. 9, no. 9, pp. 1416–1423, 2014.