National Research University Higher School of Economics

as a manuscript

Sokolova Anna

Neural network model for human recognition by gait in different types of video

PhD Dissertation Summary

for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Moscow — 2020

This work was prepared at National Research University Higher School of Economics.

Academic Supervisor: Anton S. Konushin, PhD, Associate Professor, National Research University Higher School of Economics

Dissertation subject

According to Maslow studies [1], safety need is one of the basic and fundamental human needs. People tend to protect themselves, preserve their housing from illegal invasion and save property from stealing.

With the development of modern video surveillance systems, it becomes possible to capture everything that happens in a certain area and then analyze the obtained data. Using video recordings, one can track the movements of people, determine illegal entry into private territory, identify criminals who get captured by the cameras, control access to restricted objects. For example, video surveillance systems help to catch burglars, robbers or arsonists, automatically count the number of people in a line or in crowd, and analyze the character of their movements reducing the amount of subjective human intervention and decreasing the time required for data processing. Besides this, being embedded in the currently widely used home assistance systems ("smart home"), they can distinguish family members and change behavior depending on the personality. For example, it can be configured to conduct different actions if it captures a child or elderly person.

Dissertation topic and its relevance

Recently, the problem of recognizing a person in a video (see Fig. 1) has become particularly urgent. A human's personality is identifiable in a video based on several criteria, and the most accurate one is facial features. However, the current recognition quality allows to entrust decision-making to a machine only in a cooperative mode, when person's face is compared with a high quality photograph in a passport. In real life (especially when committing crimes), a person's face may be hidden or poorly visible due to bad view, insufficient lighting, or the presence of a headgear, mask, makeup, etc. In this case, another characteristic is required to make the recognition, and gait is the possible one. According to biometric and physiological studies [2; 3], the manner the person walks is individual and can not be falsified, which makes gait a unique identifier comparable to fingerprints or the iris of the eyes. In addition, unlike these "classic" methods of identification, gait can be observed at a great distance, it does not require perfect resolution of the video and, most importantly, no direct cooperation with a human is needed, thus, a human may not know that he is being captured and analyzed. So, in some cases gait serves as the only possible sign for determining a person in the video surveillance data.

The gait recognition problem is very specific due to the presence of many factors that change the gait visually (different shoes; carried objects; clothes that hide parts of the body or constrain movements) or affect the internal representation of the gait model (angle, various camera settings). In this regard, the quality and reliability of identification by gait is much lower than by face, and, despite the success of modern methods of computer vision, this problem

Figure 1 — Informal problem statement: having a video with a walking person one needs to determine this person's identity from the database

has not been solved yet. Many methods are applicable solely to the conditions present in the databases on which they are trained, which limits their usability in real life.

In addition to the classic surveillance cameras that store everything that happens in the frame $25 - 30$ times per second, other types of sensors, in particular, dynamic vision sensors (Dynamic Vision Sensors, DVS [4—6]), are gaining popularity in recent years. Unlike conventional video cameras, the sensor, like the retina, captures changes in intensity in each pixel, ignoring points with constant brightness. Under the conditions of a static sensor, events at background points are very rarely generated, preventing the storage of redundant information. At the same time, the intensity at each point is measured several thousand times per second, which leads to the asynchronous capture of all important changes. As a result, such a stream of events turns out to be very informative and suitable for obtaining the data necessary for solving many video analysis tasks that require the extraction of dynamic characteristics, including gait recognition.

Dynamic vision sensors are now a promising rapidly developing technology, which leads to the need to solve video analysis tasks for the received data. Despite the constant development of computer vision methods, no approaches to solving the gait recognition problem according to the data of dynamic vision sensors have yet been proposed, and it represents a vast field for research.

The methods of deep learning based on the training of neural networks have become the most successful in solving computer vision problems in recent years. Attributes taught by neural networks often have a higher level of abstraction, which is necessary for high-quality recognition. This allows to achieve outstanding results in solving such problems as the classification of video and images, image segmentation, object detection, visual tracking, etc. However, despite the success of deep learning methods, classical computer vision methods are still ahead of neural networks in some gait recognition benchmarks and both approaches have not achieved acceptable accuracy for full integration yet.

Goals and objectives of the research

This research aims to develop and implement the neural network algorithm for human identification in video based on the motion of the points of human figure that is stable to viewing angle changes, different clothing and carrying conditions. To achieve this goal the following tasks are set:

1. Develop and implement the algorithm for human identification in video analysing the optical flow.
2. Develop an implement the multiview algorithm for gait recogntion based on the analysis of the motion of different human body parts.
3. Develop the algorithm for human recognition in the event-based data from Dynamic Vission Sensors.

Formal problem statement

The formal research objects are video surveillance frame sequences v and event streams e from the dynamic vision sensors where the moving person is captured. Having the labelled gallery D given one needs to deternime, which person from the gallery appears in video, i.e. the identity of the person in video is to be defined.

Let the gallery be defined as

$$D = \{(v_i, x_i)\}_{i=1}^{N},\ x_i \in P,$$

where N is the number of sequences and P is the set of subjects. The label of the subject $x \in P$ is to be found for the video v under investigation. The goal is to construct some similarity measure S according to which the closest object will be searched for in the gallery.

$$S(v,v_i) \rightarrow \min_{v_i:\ \exists x_i:\ (v_i,x_i)\in D}$$

The problem for the event streams is stated similarly.

A set of restrictions is imposed on all the sequences:

- each video in the gallery contains exactly one person;
- each person is captured full length;
- no occlusions;
- static camera;
- the set of posible camera settings (its height and tilt) is limited.

The described conditions are introduced due to the limitations of the existing datasets and benchmarks.

Novelty and summary of the authors main results

In this thesis, the author introduces the original method for human recognition by gait stable to view changes, reducing the length of the video sequences and dataset transfer. The following is the list of the main research results. The list of the corresponding publications can be found in section **Publications** at the page 7.

1. Side-view gait recognition method which analyses the points translations between consecutive video frames is proposed and implemented. The method shows state-of-the-art quality on the side-view gait recognition benchmark.
2. Multi-view gait recognition method based on the consideration of movements of the points in different areas of human body is proposed and implemented. The state-of-the-art recognition accuracy is achieved for cerrain viewing angles and the best at the investigation time approaches are outperformed in verification mode.
3. The influence of the point movements in different body parts on the recognition quality is revealed.
4. Gait recongition method stable to dataset transfer is proposed and implemented.
5. Two approaches for view resistance improvement are proposed and implemented. Both methods increase the cross-view recognition quality and complement each other being applied simultaneously. The model obtained using these approaches outperforms the state-of-the-art methods on multi-view gait recogntion benchmark.
6. The method for human recognition by motion in the event-based data from dynamic vision sensor is proposed and implemented. The quality close to conventional video recognition is obtained.

The described results are original and obtained for the first time. Below, the author's contributions are summarized in four main points.

1. The first gait recognition method based on the investigation of the point movements in different parts of the body is proposed and implemented.
2. Two original methods of view resistance improvements are proposed. In these approaches the auxiliary model regularization is made and the descriptors are projected into the special feature space decreasing the view dependency.
3. The original research of the gait recognition algorithm transfer between different data collections is made.
4. The first method for human recognition by motion in the event-based data from dynamic vision sensor is propose and implemented.

Practical significance

Gait recognition is an applied problem of computer vision. Being proposed according to natural and mathemathical reasons, all the suggested methods and approaches aim to be applicable. Thus, being implemented, the proposed human identification methods can be intergrated to different automation systems. For example, the developed approach can be used in the home assistance systems ("smart home") which recognize the family members and

change the behaviour depending on the captured person. Being united with the alarm, the system can respond to the appearance of the people not included to the family, and track the illegal entrance into the private houses.

Besides this, the gait identification algorithm can be used in crowded places, such as train stations and airports, where it is not possible to take close-up shots, but there is an obvious need to track and control access.

Publications and approbation of the research

Main results of this thesis are published in the following papers. The PhD candidate is the main author in all of these articles.

First-tier publications:

- A. Sokolova, A. Konushin, Pose-based deep gait recognition // IET Biometrics, 2019, (Scopus, Q2).

Second-tier publications:

- A. Sokolova, A. Konushin, Gait recognition based on convolutional neural networks // International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, 2017 (Scopus).
- A. Sokolova, A. Konushin, Methods of gait recognition in video // Programming and Computer Software, 2019 (WoS, Scopus, Q3).
- A. Sokolova, A. Konushin, View Resistant Gait Recognition // ACM International Conference Proceeding Series, 2019 (Scopus).

The results of this thesis have been reported at the following **conferences and workshops**:

- ISPRS International workshop "Photogrammetric and computer vision techniques for video surveillance, biometrics and biomedicine" – PSBB17, Moscow, Russia, May 15 – 17, 2017. Talk: Gait recognition based on convolutional neural networks.
- Computer Vision and Deep Learning summit "Machines Can See", Moscow, Russia, June 9, 2017. Poster: Gait recognition based on convolutional neural networks.
- 28th International Conference on Computer Graphics and Vision "GraphiCon 2018", Tomsk, Russia, September 24 – 27, 2018. Talk: Review of video gait recognition methods.
- Samsung Biometric Workshop, Moscow, Russia, April 11, 2019. Talk: Human identification by gait in RGB and event-based data.
- 16th International Conference on Machine Vision Applications (MVA), Tokyo, Japan, May 27 – 31, 2019. Poster: Human identification by gait from event-based camera.
- 3rd International Conference on Video and Image Processing (ICVIP), Shanghai, China, December 20 – 23, 2019. Talk: View Resistant Gait Recognition (best presentation award).

Contents of the work

The **Introduction** describes the relevance of the disseration topic and formal statement of the investigated problem. The goals and objectives are formulated, as well as scientific novelty and practical significance of the work.

It is followed by the **fist chapter** devoted to the background of the problem: details and difficulties of the task, decription of the existing data collections and the review of modern methods aiming at the problem solution. The advantages and disadvantages of these methods are presented in order to justify the relevance of this research. Three most popular and general data collections ("TUM Gait from Audio, Image and Depth" (TUM-GAID) [7], CASIA Gait Dataset B [8] and OU-ISIR Large Population Dataset (OULP) [9]) are selected for the experiments and comparison of the proposed method with state-of-the-art ones.

In the **second chapter** the baseline algorithm for gait recognition problem from the side view is proposed. The chapter starts with the consideration of adjacent computer vision problems having a lot in common with gait recognition and motivation of optical flow usage for motion-based identification. After that, the formal pipeline of the proposed algorithm is provided (see Fig. 2) consisting of the following steps:

1. Data preprocessing;
2. Neural network backbone;
3. Feature postprocessing;
4. Classification.

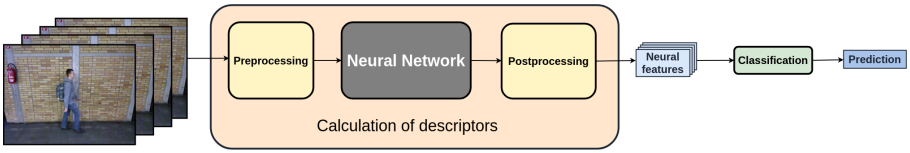


Figure 2 — General algorithm pipeline

The following subsections focus on the steps enumerated above.

The subsection 2.2.1 decribes data preprocessing algorithm. Optical flow is proposed to be the main sourse of information for feature extraction. Reflecting the points movements between the pairs of the consecutive frames optical flow is a direct motion characteristics. Being applied together with the features obtained from the raw color frames it has shown a high quality in the action recognition problem [10]. It gives hope to suggest that such data is really enough for successful analysis of video.

In order to enhance the temporal component and to use not only instantaneous but continuous motion, the maps are used not separately, but jointly. Several consecutive maps are stacked into one block combining short-term and long-term dynamic characteristics of gait.

The whole preprocessing algorithm has the following structure:

1. Human figure is detected in each frame.
2. Optical flow maps are calculated between the pairs of consecutive frames.
3. The maps are stacked into the 10 frame blocks with 5 frame intersection.
4. The bounding box containing the human figure in each frame is computed for each block. The rectangle is cropped from the block according to this bounding block.

The subsection 2.2.2 is devoted to the neural network backbone of the model. Firstly, the data augmentation procedure is described. It is required for neural network training due to relatively small size of all the available gait collections. Then, the neural architectures and training process are considered. The classical CNN [11] which was the first to show high recognition quality, was selected as the baseline network. However, a lot of different approaches to neural network construction have been proposed recently. Beside the baseline model, the fusion of several streams of Siamese network is investigated as well as a VGG-like [12] model.

In all cases, the network is trained by stochastic gradient descent (SGD) for classification task, getting the $(2N \times W \times H)$ tensor and outputting the vector of probability distribution over the classes which dimansionality equals the number of training subjects. The network is being trained once on the fixed dataset and can further be used for the feautre extraction. The top dense layer is being removed and the outputs of the last hidden layer are used as gait representations. Such an approach allows to use one model without re-training even when the new subjects are added to the database. In this case, only the feature classificational model should be changed which requires much less time than CNN fine-tuning.

In the subsection 2.2.3, the extracted feature postprocessing is described. Although the features already have low dimensionality, its further reduction is proposed. The application of principal components analysis (PCA) helps to get rid of the noise and increase the recognition accuracy. Besides this, such a low-dimensional projection accelerates the fitting of any classifier making the whole algorithm more effective.

However, the PCA decomposition is not the only proposed modification of the descriptors. The further recognition improvement can be made by Triplet Probability Embedding (TPE) [13]. This method aims to find a low-dimensional discriminative 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

$$S_W(v_a, v_p) > S_W(v_a, v_n), \tag{1}$$

where $S = S_W$ is a 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 belongs to the other one. The problem of finding the projection matrix W can be reduced to triplet loss optimization solved by gradient descent.

For the gait recognition problem, this method has been transformed as follows. Instead of cosine similarity measure and the scalar product, the Euclidian distance is used, and the sign in the inequality 1 is reversed. Inroducing the probability p_{apn} of the triplet (v_a, v_p, v_n)

$$p_{apn} = \frac{e^{S_W(v_a, v_p)}}{e^{S_W(v_a, v_p)} + e^{S_W(v_a, v_n)}} \tag{2}$$

the problem can be reduced to likelihood maximization which is solved by SGD using hard negative mining.

Applying the described embedding and averaging the descriptors over all the blocks, the video gait features are obtained. Having such descriptors computed for each video, the closest one from the database is found and returned as the response. Additionally to the "classical" Euclidian distance between vectors, Manhattan distance is considered as the base for nearest neighbour (NN) classifier. It turns out that this metrics is more stable and appropriate for measuring the similarity of gait descriptors and shows a better result in the majority of experiments.

The section 2.3 is devoted to the details of the experiments and their numerical evaluation. The first set of the experiments was conducted to evaluate the efficiency of proposed approach in general. A network has been trained from scratch with the top layer of size equal the number of training subject. Then the low-dimensional embedding was trained and the index for the search was constructed on the fitting data. The results of the experiments are shown in Table 1. The best results are obtained by VGG-like network which was the most successfull at the time of the study.

Besides this, the experiments on the algorithm transfer between the datasets are provided in this section. Due to the great variability of the gait and absence of one general gait collection covering all the variations, the models suffer from overfitting on the conditions presented in training dataset. The original experiment on model transfer and joint training has been conducted and described in this section. These experiments were made using truncated VGG which is the best of considered architectures, and the NN classifier based on L_1 metrics without any extra embedding. Table 2 shows the accuracy of such transfer classification. The training process is the same in each model, but training and testing datasets are different. This table confirms that joint training on two datasets significantly increases the recognition quality, which shows that being visually similar, the collections distinguish leading to model overfitting.

Table 1 — Results on TUM-GAID dataset

Method			Evaluation	
Architecture	Embedding (dimensionality)	Metrics	Rank-1 [%]	Rank-5 [%]
CNN-M	PCA (1100)	L_1	93,22	98,06
CNN-M	PCA (1100)	L_2	92,79	98,06
CNN-M	PCA (600)	L_1	93,54	98,38
CNN-M	TPE (450)	L_2	94,51	98,70
CNN-M fusion	PCA (160)	L_1	93,97	98,06
CNN-M fusion	PCA (160)	L_2	94,40	98,06
CNN-M fusion	TPE (160)	L_1	94,07	98,27
CNN-M fusion	TPE (160)	L_2	95,04	98,06
VGG	PCA (1024)	L_1	97,20	99,78
VGG	PCA (1024)	L_2	96,34	99,67
VGG	TPE (800)	L_1	**97,52**	**99,89**
VGG	TPE (800)	L_2	96,55	99,78

Table 2 — The quality of transfer learning

Training Set	Testing Set	
	CASIA	TUM
CASIA	74,93%	67,41%
TUM	58,20%	97,20%
CASIA + TUM union	72,06%	96,45%

The chapter ends with the conclusion section summarizing the results of the investigation.

The **third chapter** focuses on the developement of the baseline method. Continuing using optical flow as the main source of information, it is suggested that the algorithm should consider the motion around some body parts in more details in order to catch more important gait characteristics. The pipeline of such approach is shown in Fig. 3. The chapter starts with the motivation section where the need to consider different parts of the body additionally to the full body is explained.

Section 3.2 focuses on the proposed algorithm. This section consists of several subsections reflecting the steps of the algorithm. Firstly, the body

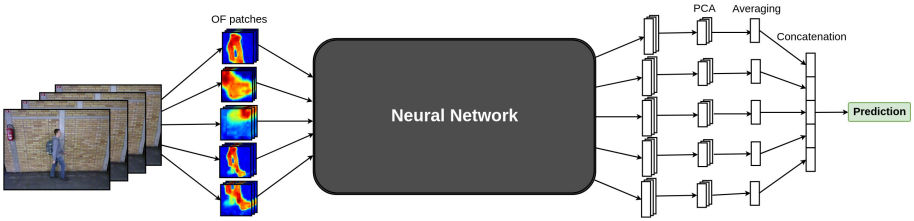


Figure 3 — The pipeline of the pose-based gait recognition algorithm

part selection process is described and the choise is justified. The set of the considered areas contains the body parts of different sizes from the full body to small areas around the joints, which allows to obtain the features from different scales and pay more attention to some areas of the human figure. The upper joints of the body (such as head or hands) are supposed to be less informative for gait recognition since they can move independently during the walk, thus, the five element set of body parts is considered: full body, upper and lower halves of the body and two legs.

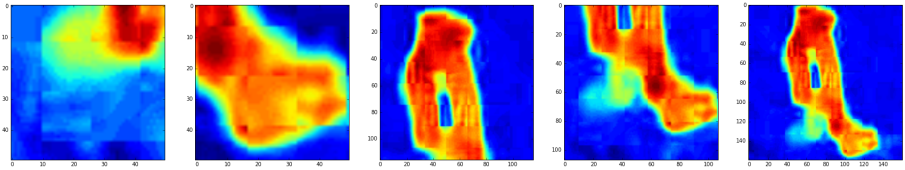


Figure 4 — The visualizations of optical flow maps in five areas of the human figure

Then, the subsection 3.2.2 is provided which is devoted to the main "body" of the algorithm, neural network pipeline. Similarly to the baseline model, the optical flow maps are inputted into the network, but several areas are cropped from each map corresponding to the selected body parts (see Fig. 4). For all the parts of the body the same convolutional neural network is trained to predict the probability distribution over classes for distinct optical flow patches. The architecture influence on the recognition quality is investigated as well as in chapter 2, but the VGG-like backbone is chosen as the baseline since it has shown the best results in the side-view approach. However, the WideResNet model based on residual connections turns out to be the most successful.

The section ends with the subsection 3.2.3 describing the feature aggregation method and final classification. Since several patches are cropped from each frame, the same number of features are extracted. The naive aggregation method by averaging the descriptors over both time and body surprisingly achieves high accuracy, however the better results can be obtained by the following procedure. The features are averaged over time and then the

mean video descriptors for each body part are concatenated into one high-dimensional vector. Such an approach does not mix the components of the figure and allows to save the information extracted from different scales and areas.

The next section of the chapter provides the experimental evaluation of the approach. The models based on different neural architectures, set of body parts and aggregation methods are considered and compared. The best quality is obtained by Wide Residual Network and temporal averaging. The dependence of the identification accuracy on the set of body parts is presented in Table 3. According to the experiments, the addition of small parts of the body really

Table 3 — Recognition quality of models based on different sets of body parts on TUM-GAID dataset

Body parts	Rank-1 [%]	Rank-5 [%]
legs	79,7	86,5
upper body	96,2	99,7
lower body	96,3	99,6
full body	98,9	**100,0**
full body, upper body, lower body	99,4	**100,0**
full body, upper body, lower body, legs	**99,8**	99,9

increases the recognition quality.

Besides this, the experimental evaluation of multi-view recognition is provided in this section. The main advantage of this approach is the view-independence: it can be applied to the data captured under any view. The Table 4 reflects the average cross-view recogniiton accuracy and its comparison with the state-of-the-art methods [14; 15].

Table 4 — Average recognition rates for three angles on CASIA database

Model	Average Rank-1 [%] Probe View			
	54°	90°	126°	Mean
WideResNet + PCA (230), concat, L_1	**77,8**	**68,8**	74,7	**73,8**
SPAE [14]	63,3	62,1	66,3	63,9
Wu [15]	**77,8**	64,9	**76,1**	72,9

To compare the algorithm with two more state-of-the-art cross-view recognition methods [15—17], the experiments on OULP collection were conducted. The results of one of these experiments and accuracy comparison are presented in Table 5.

Table 5 — Comparison of *Rank-1* on OULP dataset

Model	Rank-1 [%] Angular difference				
	0°	10°	20°	30°	Mean
Wide ResNet, L_1	98,4	98,2	97,1	94,1	97,0
Takemura [16]	99,2	99,2	98,6	97,0	98,8
Wu [15]	98,9	95,5	92,4	85,3	94,3
LDA [17]	97,8	97,1	93,4	82,9	94,6

One more set of experiments focuses on the dependency of recognition quality on the length of the video. Although one gait cycle lasts about 1 second, which is much shorter than the length of video sequence, it turns out that the longer the considered video is, the better it is recognized. The expanded time of analysis is required because body point motions are similar but not identical for each step and, correspondingly, using long sequences makes recognition more stable to small inter-step changes in walking style.

The chapter ends up with the conclusion section, where the advantages and weak points of the approach are summarized.

The main drawback of the proposed model, similarly to all the other existing approaches is low quality of cross-view recognition. This challenge is the most complex in gait recognition and it is faced in the **fourth chapter** devoted to the increasing of view stability. Considering Nearest Neighbour method as the final classifier of gait descriptors, one needs to make the descriptors of the same subject as close to each other as possible regardless the viewing angle the video had been captured under. At the same time, the vectors corresponding to different subjects should be far from each other in order the spatially close vectors to have the same labels. According to these reasonings, the features trained by neural network should be given a cluster structure. Two techniques are proposed for this purpose.

The first approach consists in the modification of training process in order to make the neural features view independent. It is proposed to add an auxiliary loss function to penalize the difference of person's gait features under the fixed view from the average feature over all 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. Let $d^{i,\alpha}$ be the average of last hidden layer outputs of the first batch, and d^i be the average hidden output of the second batch. Then the loss function equals

$$L_{reg} = \lambda L_{view} + L_{clf}, \tag{3}$$

where $L_{view} = ||d^{i,\alpha} - d^i||^2$ is the described view loss corresponding to optimization task, $L_{clf} = LogLoss(\{b^{i,\alpha}\}) + LogLoss(\{b^i\})$ is conventional

cross-entropy for classification applied to both batches, and λ is relative weight of view term in total loss. The view term added to conventional loss function can be considered as the regularization to prevent view memorizing. The full training process alternating the steps with and without regularization is described in this section. It is shown that such a transformed loss function leads to great decrease of the inregularized loss.

The next section of this chapter describes the second technique of view dependence overcoming. It is suggested that the features can be embedded in a new subspace after the extraction from the network. The cross-view triplet probability embedding is trained to make the features of the same subject obtained under different angles closer to each other, and the features of different subjects further from each other even if the views coincide. Such an embedding is a modification of TPE method [13], but the addition of view dependency while triplet sampling significantly influences the multi-view recognition quality and improves it more than the classical embedding. The scheme of the view resistant gait recognition method is shown in Fig. 5.

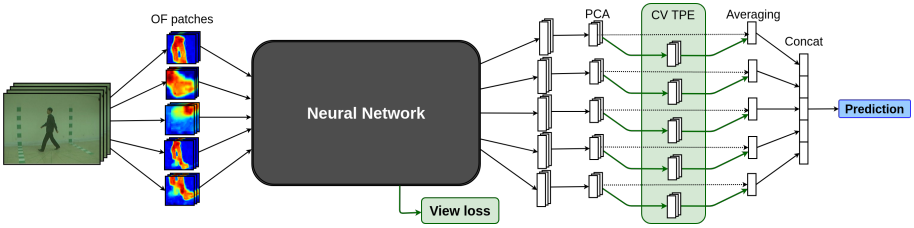


Figure 5 — The pipeline of gait recognition algorithm with two modifications (view loss and CV TPE) shown in green boxes

The description of the approaches is followed by the experimental evaluation section. It is shown (see Tab. 6) that joint usage of two proposed techniques affects the recognition more than their separate application. It confirmes that they really complement each other and increase the view stability of the model. The comparison of all the models with state-of-the-art approaches [14; 15; 18] presented in Table 6 demonstrates the superiority of the combination of the developed approaches over all the models existed at the time of the study.

The last **fifth chapter** is devoted to the investigation of human recognizability in the event stream and applicability of the developed gait recognition methods to such stream. The pipeline of the processing algorithm is illustarated in Figure 6. The chapter begins with the detailed description of dynamic visual sensors and justification of the need of the problem solution in the event-based data. Due to the effectiveness and sensitivity of DVS, they become wide spread in many fields and often substitute the conventional

Table 6 — Comparison of average cross-view recognition accuracy of proposed approaches and state-of-the-art models

Method	Average accuracy [%] Probe angle			
	0°	54°	90°	126°
Part-based model	59,7	80,1	68,9	77,7
Part-based model + View Loss	65,9	83,6	74,6	83,7
Part-based model + TPE	56,9	82,6	73,5	82,4
Part-based model + CV TPE	62,6	84,6	75,6	84,4
Part-based model + View Loss + CV TPE	**69,3**	86,3	**75,8**	**86,0**
SPAE [14]	-	63,3	62,1	66,3
Wu [15]	54,8	77,8	64,9	76,1
GaitSet [18]	64,6	**86,5**	75,5	**86,0**

cameras in video surveillance systems, leading to the need of developement of computer vision methods for dynamic vision data.

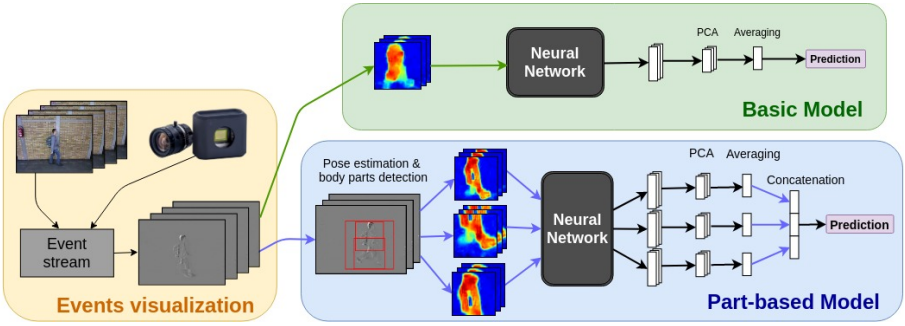


Figure 6 — The pipeline of the event-based gait recognition algorithm

The second section of the chapter focuses on the data visualization. Since the gait recognition is a human motion analysis problem, the mutual arrangement of the points and their relative changes are of much interest rather than discrete events in distinct points. Besides this, in order to apply convolutional neural network methods of video analysis, the set of events in distinct points should be transformed into two- or three-dimensional tensors (depending on the channels number), containing all the captured data.

The simplest way of data visualization which, however, allows to achieve high recognition quality, consists in the summarizing the events in each point in a temporal window of a certain length (see Fig. 7). The sum is calculated taking the event polarity into account (1 or -1 depending on increasing or

decreasing of pixel intensity). Such visualized frames are well interpretable: if nothing has happened in the point during time interval, the resulting value equals zero, and the points in which some changes occur get non-zero value depending on the number of events captured in the point during the time interval and their polarity. Thus, these visualizations can be considered as the measure of happenings in each point. Although these constructed images are not real frames and many appearance characteristics get lost, the moving human figure is visually recognizable in these grey-scale images and this figure is the only moving object in the frame due to stationarity of the background and chosen aggregation method.

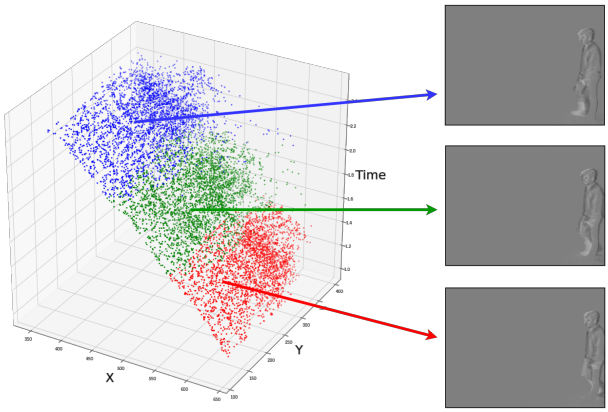


Figure 7 — The transformation of the event stream into a sequence of frames

The third section of the chapter describes the changes that should be made in the proposed approach to apply it to event visualizations. These changes concern figure detection, pose estimation and body part selection procedures. The detection problem becomes simpler, since in case of static camera and no occlusions, the human figure turns out to be the only object in some area leading to event generation (see the rigth part of Figure 7). Thus, the "semantical" detection is not needed and can be reduced to the separation of a non-uniform human from the background that is almost monophonic. Thereby, a basic blob model was constructed which considers the average relative deviation of pixel brightness in each row and column of visualized stream from the median value and detects when these deviations exceed some threshold.

Unlike detection, the pose estimation problem becomes more complex in case of event-based data. Due to the visual differences, the state-of-the-art methods trained for pose estimation in conventional RGB images do not cope with event-based ones without additional fine-tuning and require re-training. An important difference between grey-scale event visualizations and

conventional images is the leg invisibility. Being on the ground the supporting leg does not move and, thus, no events are generated for a while in this area and the corresponding leg becomes invisible (Figure 8). And legs not only influence on the boundaries of the boxes, but completely define some of the boxes. For these reasons, the reduced set is considered containing only three big parts corresponding to upper and lower body parts and the full body.

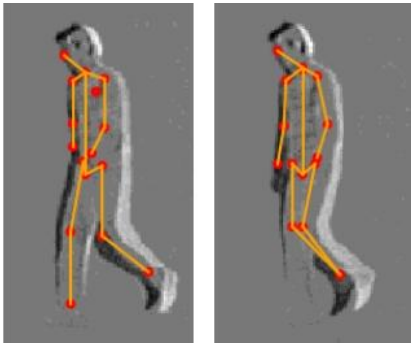


Figure 8 — Examples of estimated skeletons with invisible leg

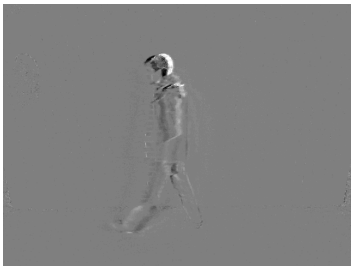
The fourth section concerns the event-based data. The author is the first to investigate the gait recognizability in DVS data, so, there are no publicly available data collections appropriate for the problem solution. Thus, the event stream has to be simulated. Since the gait is a special motion, it is hardly generatable from scratch, the initialization is required, thus it is proposed to simulate streams based on the existing gait collections. The data generation has been conducted in two steps: the intermediate frame interpolation and event generation itself. The first step is required since the conventional video sequences have low frequency in contrast to event streams. The simplest way of interpolation is pixel-wise linear approximation, which leads to a good result visually similar to real data. The further event generation is made by comparison the pixel intensity change with some threshold. Figure 9 shows the visualizations of real and simulated data confirming their similarity.

After the description of the algorithm, the experimental section is provided. The models based on different body parts are evaluated and compared, and it is proven that the legs neighborhoods are really uninformative in this model and their usage worsens the quality of identification. Table 7 presents the numerical comparison result and shows that the event-based model is only 0,8% behind the original RGB model.

In the **conclusion** the main results of the thesis are presented:

1. Side-view gait recognition method which analyses the points translations between consecutive video frames is proposed and implemented. The method shows state-of-the-art quality on the side-view gait recognition benchmark.

real data



simulated data

Figure 9 — The visualizations of the real and simulated event streams

Table 7 — The quality of the end-to-end model on simulated TUM dataset

Body part set	Rank-1 [%]
Basic full body	98,0
Pose-based, three parts	99,0
Pose-based, five parts	98,8
Original RGB pose-based model	99,8

2. The first multi-view gait recognition method based on the consideration of movements of the points in different areas of human body is proposed and implemented. The state-of-the-art recognition accuracy is achieved for certain viewing angles and the best at the investigation time approaches are outperformed in verification mode.
3. The influence of the point movements in different body parts on the recognition quality is revealed.
4. The original research of the gait recognition algorithm transfer between different data collections is made.
5. Two original approaches for view resistance improvement are proposed and implemented. Both methods increase the cross-view recognition quality and complement each other being applied simultaneously. The model obtained using these approaches outperforms the state-of-the-art methods on multi-view gait recogntion benchmark.
6. The first method for human recognition by motion in the event-based data from dynamic vision sensor is proposed and implemented. The quality close to conventional video recognition is obtained.

References

1. *Maslow, A.* Motivation and Personality / A. Maslow. — Oxford, 1954. — 411 p.
2. *Cutting, J. E.* Recognizing friends by their walk: Gait perception without familiarity cues / J. E. Cutting, L. T. Kozlowski // Bulletin of the Psychonomic Society. — 1977. — Vol. 9, no. 5. — P. 353–356.
3. *Murray, M. P.* GAIT AS A TOTAL PATTERN OF MOVEMENT / M. P. Murray // American Journal of Physical Medicine & Rehabilitation. — 1967. — Vol. 46.
4. *Lichtsteiner, P.* A 128× 128 120 dB 15 μs Latency Asynchronous Temporal Contrast Vision Sensor / P. Lichtsteiner, C. Posch, T. Delbruck // IEEE Journal of Solid-State Circuits. — 2008. — Vol. 43, no. 2. — P. 566–576.
5. 4.1 A 640×480 dynamic vision sensor with a 9µm pixel and 300Meps address-event representation / B. Son [et al.] // 2017 IEEE International Solid-State Circuits Conference (ISSCC). — 02/2017. — P. 66–67.
6. A 240 × 180 130 dB 3 µs Latency Global Shutter Spatiotemporal Vision Sensor / C. Brandli [et al.] // IEEE Journal of Solid-State Circuits. — 2014. — Oct. — Vol. 49, no. 10. — P. 2333–2341.
7. The TUM Gait from Audio, Image and Depth (GAID) database: Multimodal recognition of subjects and traits. / M. Hofmann [et al.] // J. of Visual Com. and Image Repres. — 2014. — Vol. 25(1). — P. 195–206.
8. *Yu, S.* A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition. / S. Yu, D. Tan, T. Tan // Proc. of the 18'th ICPR. Vol. 4. — 2006. — P. 441–444.
9. The OU-ISIR Gait Database Comprising the Large Population Dataset and Performance Evaluation of Gait Recognition / H. Iwama [et al.] // IEEE Trans. on Information Forensics and Security. — 2012. — Oct. — Vol. 7, Issue 5. — P. 1511–1521.
10. *Simonyan, K.* Two-stream convolutional networks for action recognition in videos. / K. Simonyan, A. Zisserman // NIPS. — 2014. — P. 568–576.
11. Return of the devil in the details: Delving deep into convolutional nets. / K. Chatfield [et al.] // Proc. BMVC. — 2014.
12. *Simonyan, K.* Very deep convolutional networks for large-scale image recognition. / K. Simonyan, A. Zisserman // ICLR. — 2015.
13. Triplet probabilistic embedding for face verification and clustering / S. Sankaranarayanan [et al.] // 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS). — 2016. — P. 1–8.

14. Invariant feature extraction for gait recognition using only one uniform model / S. Yu [et al.] // Neurocomputing. — 2017. — Vol. 239. — P. 81—93.

15. A Comprehensive Study on Cross-View Gait Based Human Identification with Deep CNNs / Z. Wu [et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2016. — Mar. — Vol. 39.

16. On Input/Output Architectures for Convolutional Neural Network-Based Cross-View Gait Recognition / N. Takemura [et al.] // IEEE Transactions on Circuits and Systems for Video Technology. — 2019. — Sept. — Vol. 29, no. 9. — P. 2708—2719.

17. *Belhumeur*, *P. N.* Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection / P. N. Belhumeur, J. P. Hespanha, D. J. Kriegman // IEEE Trans. Pattern Anal. Mach. Intell. — 1997. — July. — Vol. 19, no. 7. — P. 711—720.

18. GaitSet: Regarding Gait as a Set for Cross-View Gait Recognition / H. Chao [et al.] // AAAI. — 2019.