Master Thesis

Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Department of Cybernetics

Long-Term Person Re-Identification In Video

Martina Tichá

Supervisor: Ing. Jiří Čermák, Ph.D.
Field of study: Open Informatics
Subfield: Computer vision and digital image processing
December 2019

CTU
CZECH TECHNICAL
UNIVERSITY
IN PRAGUE

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Tichá Martina** Personal ID number: **475409**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Cybernetics**

Study program: **Open Informatics**

Branch of study: **Computer Vision and Image Processing**

II. Master’s thesis details

Master’s thesis title in English:

Long Term Person Reidentification in Video

Master’s thesis title in Czech:

Dlouhodobá reidentifikace lidí ve videu

Guidelines:

Long term person reidentification in video streams is a challenging task, since the person‘s clothes may change and the availability of facial data is unreliable due to small resolution or bad viewing angle. Solving such task would have a large impact in security, law enforcement and other areas.
The student will study the state-of-the-art approaches for person reidentification in the literature. Based on this analysis, she will devise and implement an approach suitable for long-term person reidentification. The performance of the approach will be empirically evaluated on one of the available datasets for person reidentification purposes (e.g., MARS Dataset).

Bibliography / sources:

[1] Peng Zhang, Qiang Wu, Jingsong Xu, Jian Zhang - Long-term Person Re-identification using True Motion from Videos - 2018 IEEE Winter Conference on Applications of Computer Vision
[2] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, Qi Tian - MARS: A Video Benchmark for Large-Scale Person Re-identification - Peking Universit 2016
[3] Md Jan Nordin, Ali Saadoon - A Survey of Gait Recognition Based on Skeleton Model for Human Identification - Research Journal of Applied Sciences, Engineering and Technology, 2016

Name and workplace of master’s thesis supervisor:

Ing. Jiří Čermák, Ph.D., Blindspot Solutions, Prague

Name and workplace of second master’s thesis supervisor or consultant:

doc. Ing. Jiří Vokřínek, Ph.D., Department of Computer Science, FEE

Date of master’s thesis assignment: **05.06.2019** Deadline for master's thesis submission: **07.01.2020**

Assignment valid until: **19.02.2021**

Ing. Jiří Čermák, Ph.D. Supervisor’s signature	doc. Ing. Tomáš Svoboda, Ph.D. Head of department’s signature	prof. Ing. Pavel Ripka, CSc. Dean’s signature

III. Assignment receipt

The student acknowledges that the master’s thesis is an individual work. The student must produce her thesis without the assistance of others, with the exception of provided consultations. Within the master’s thesis, the author must state the names of consultants and include a list of references.

.
Date of assignment receipt Student’s signature

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Ing. Jiří Čermák, Ph.D., who guided me throughout writing this thesis. I would also like to thank all the great people at Blindspot.ai, who have been supporting me for the whole time of my studies at the Czech Technical University.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of the university theses.

. . . .

Martina Tichá
Prague, 16th December 2019

Abstract

The task of person re-identification in a video has been a subject of research for many years, since the information that is usually used for person identification, such as face or fingerprints, are not reliably available. The current approaches for the person re-identification usually rely on visual features such as clothing or hairstyle and so they assume that the recorded identities have not changed their appearance significantly. This limits their usage to a short-term event. The goal of this thesis is to study approaches, which do not rely on visual features and can possibly be used in a long-term scenario.

Keywords: person re-identification, artificial neural network, edge detection, skeleton extraction, computer vision

Supervisor: Ing. Jiří Čermák, Ph.D.
Blindspot Solutions, Praha

Abstrakt

Úloha reidentifikace osob z videa je předmětem výzkumu po mnoho let, jelikož informace, které jsou běžně používány k identifikaci lidí, nejsou vždy k dispozici. Současné přístupy k reidentifikaci osob jsou většinou zaměřeny na aspekty typu oblečení, či účes, čímž předpokládají, že vzhled pozorovaných osob se nijak významně nezměnil. Tato skutečnost omezuje reidentifikaci osob na krátkodobou událost. Cílem této práce je zkoumat přístupy, které nejsou založeny na vzhledu pozorovaných osob a díky tomu mohou být použity ke dlouhodobé reidentifikaci osob.

Klíčová slova: rozpoznávání osob, umělá inteligence, detekce hran, detekce kostry, počítačové vidění

Překlad názvu: Dlouhodobá reidentifikace lidí ve videu

Contents

Figures

Tables

Chapter 1

Introduction

The task of Person Re-Identification in a video (person Re-Id) aims to re-identify a person that has already been recorded by a camera. It has been a subject of research for many years, since information usually used for identifying a person (mainly its face) is not reliably available, due to the bad viewing angle or the insufficient resolution of the camera.

There is a vast variety of practical applications of person Re-Id systems in both commercial and law enforcement areas. Surveillance systems are nowadays widely deployed for public safety. Hence, there is a high demand for technologies that can automatically detect strange behavior patterns in frequent places with high-security risks like airports, banks, or large cultural events or in areas with strong security restrictions, e.g., embassies and laboratories.

The reason why this problem is challenging is that the surveillance cameras are located in a fixed position while the recorded people are walking at various distances from the cameras. Hence, in most cases, the resolution of the video is not sufficient for recognizing details like faces. Apart from this, there are other issues like illumination changes, occlusions, pose variations, or change of clothes that make the task of person Re-Id difficult.

In the early days, the algorithms to person Re-Id were based on hand-crafted features and evaluated on small data sets. In recent years, deep learning systems utilizing newly available large-scale data sets emerged.

Most of the current approaches assume that people going across the surveillance camera have not changed their appearance significantly. This condition restricts the person Re-Id to a short-term event. However, in many places, people are likely to reappear after a long period when they have, for instance, changed their clothes. For these cases, short-term person Re-Id approaches that rely on the visual aspects of the recorded people do not suffice.

To overcome this issue, an approach that ignores the person's appearance needs to be used. One of the possibilities is to construct a model that focuses on the gait of the recorded people. To achieve this, we can first extract features describing the movement of the observed person from the raw videos and feed the model with the extracted features only. Another option is to use the original videos directly but to eliminate the visual features in the initial steps.

In this thesis, we try three approaches to eliminate the visual features in order to train model for consistent long-term Re-Id. In the first one, we transform each frame in the video sequence into contours on a white background. While this approach hides information such as the color of the clothing, some information such as the shape of the clothes the persons are wearing still remains. This can possibly confuse the model once the observed identity changes their clothes. For instance, the contour of a person in jeans is different from the one wearing a dress. In the second approach, we locate the joints of the observed people in each frame of the video sequence and use their position in time as the only input to the model. This way, all the information except for the movement characteristics of the walking people is dropped before the training. The third method is then the combination of both previous methods. We first transform each image frame into contours, and then we mark the skeletons constructed by joints connections on the silhouette of the observed identity. This way, the model is provided with both information and can itself decide which of them to use.

Chapter 2

Related Work

In this chapter, we describe the major contributions to the person Re-Id problem from the early years when person Re-Id was a part of the Person Tracking Problem, until today when person Re-Id itself gains significant attention among other computer vision problems. The main focus of this thesis is the Long-Term Person Re-Identification. However, the vast majority of the person Re-Id approaches and benchmarks so far focus on the Short-Term Person Re-Id. Since both problems are closely related, we provide an overview of methods for either of these two.

According to Zheng et al. [52], the first work, where the term "Person Re-Identification" was used, was published in 2005 by Wojciech Zajdelet et al. [48]. In this work, the authors describe a system enabling a mobile robot, equipped with a color vision system, to track people and to keep track of them when they leave the field of view. Later on, the Re-Identification task separated from Person Tracking Problem, and the focus was put on the person retrieval in multi-camera vision systems with the goal to find identical persons in footage from different cameras.

2.1 Image-based person Re-Id

The first approaches to person Re-Id focused on image matching rather than matching of the whole videos. There are two sub-tasks the Person Re-Id systems need to solve:

(a) image description - defines the feature vector that is extracted from the raw image pixels

(b) similarity metric - defines how to measure the similarity between two images

Based on how the Re-Id systems handle these two sub-tasks, we will categorize them in the same manner as in [52] into hand-crafted systems and deep-learning systems.

Hand-crafted systems

The crucial step in hand-crafted Re-Id systems is the choice of features extracted from the raw input data. The data needs to be examined, and the algorithms for feature extraction are then designed based on the properties of the data. Hence, the resulting Re-Id systems are often tuned for the input data and do not generalize well when used on a different data set.

Most commonly used features in the person description are low-level features such as color, texture, or gradient. For color representation, Prosser et al. [11] and many others use eight color channels - RGB, HS, and YCbCr. In the same work, 21 texture filters (Gabor [24] and Schmid [42]) were applied for the luminance channel. Texture features, on the other hand, can be represented as Local Binary Patterns (LBP) [38]. Low-level features can further be encoded into descriptors such as color histograms, Fisher vectors [10] or Histogram of Oriented Gradients [17]. The spacial structure of the image can be analyzed by dividing the image into a grid or stripes and the features extracted from these separated regions [38].

The second core element that needs to be chosen in hand-crafted systems is the distance metric that measures the similarity between two samples. The core idea of the similarity measurement is to keep the feature vectors of the same identity close while maintaining a distance between feature vectors of different identities. According to [52], the most commonly used distance metric is based on the Mahalanobis distance function, which is a generalization of the Euclidean distance metric that uses linear scaling and rotations of the feature spaces.

Deep-learning systems

In recent years, end-to-end Convolutional Neural Networks (CNN) have been successfully applied in many areas of computer vision [9]. The main advantage of CNNs is their ability to automatically learn the feature representation without any explicit specification of the feature extraction technique. Currently, most of the state-of-the-art Re-Id methods are based on deep-learning systems as they have significantly outperformed the previously mentioned hand-crafted systems.

The first works in Re-Id that used deep learning techniques were [16] and [47]. Since then, two main approaches to person Re-Id in deep learning emerged: classification models and verification models.

In classification models, the training data consist of labeled images categorized into classes. The input to the neural network is then a single image of an identity which needs to be classified by the model.

Verification models, on the other hand, take a set of images as their inputs and estimate the similarity between them. One example of the verification model is the Siamese neural network [26]. This network consists of two identical sub-networks whose outputs are connected by a joining neuron. The training data consists of pairs of samples which are labeled as either the same or different class. The network is then trained to recognize the data from the same class (see [52] for more details).

2.2 Video-based person re-ID

In recent years, due to the increasing data quantity, video-based person Re-Id is gaining more and more attention in research. In this case, the Re-Id algorithm consumes the whole sequence of images as its input, which enables the algorithm to collect more information about the appearance of each individual. Moreover, when keeping the input as a sequence, spatial-temporal features can be extracted for each identity. This way, the model gets additional information about movements. Similarly, as in image-based person Re-Id systems, there are two groups of video-based person Re-Id systems:

Hand-crafted systems

In 2010, Bazzani et al. [33] proposed a work where a set of images of the same identity is condensed into a highly informative signature, which is then matched with the signature of another set of images. It shows that using multiple frames per person instead of a single image significantly improves the results of the Re-Id. Such methods are usually denoted as "multi-shot person Re-Id". The disadvantage of the multi-shot strategy is that it does not incorporate any temporal cues in the model. In 2014, Wang et al. [46] proposed a method that automatically selects the most discriminative video segment from a noisy image sequence. Space-time features are then extracted from this segment. In [31], the video sequences are first decomposed into sequences of individual body units, from which Fisher vectors are extracted and combined into a descriptor of the whole video sequence. Gao at al. [12] proposes a method that uses the periodicity property of pedestrians of the "best" walking cycle from noisy motion information. To describe the video data in the selected walking cycle, the cycle is first divided into several segments. Each segment is then described by temporally aligned pooling.

Deep-learning systems

In video-based Re-Id deep-learning systems, in order to utilize the spatial-temporal features, the neural network is usually constructed so that it takes the whole sequence of images as its input. There are two prevailing approaches to achieve this. Pooling based methods aggregate appearance features (e.g., color) extracted from individual images into a vector representing the whole video sequence [36]. The main disadvantage of this method is that it cannot model well the temporal changes in human pose. The second popular method for video-based Re-Id utilizes the schema of recurrent neural networks (RNN), which are special types of neural networks that can aggregate both image-level features and human dynamics information. In [51], a special type of RNN called Long Short-Term Memory (LSTM) is fed with low-level features extracted from individual frames of input video sequences. The outputs of several such LSTMs are then connected to a softmax layer. N. McLaughlin at al. [40] uses a Convolutional Neural Network that incorporates a recurrent final layer which enables the features extracted from individual frames to be enriched by features from previously seen frames. The outputs from this neural network are then combined using temporal pooling, which provides an appearance feature of the whole sequence [40].

2.3 Long-Term Person Re-Id

All the above-mentioned methods are based on appearance matching of the observed identities, which limits their application to a Short-Term Person Re-Id. To build a model that is able to re-identify a person after a long time when visual features like clothing are no longer reliable, one needs to utilize personal characteristics that do not change in time and that are unique for each identity. To my knowledge, the only contribution to Long-Term Person Re-Id so far is in 2018 by Zhang et al. [41]. In this work, partially inspired by the success of dense trajectory on action recognition [23], the authors present a model based on a hypothesis that people keep constant motion patterns under non-distraction walking conditions. In this model, the human body is divided into several fundamental body-action primitives, where for each of them, motion descriptors are found. Fisher vectors [10] are then utilized to respectively summarize the trajectory-aligned descriptors within each body-action unite (comprised by the fundamental body-action primitives) in the predefined body-action pyramid model. This way, both local and global motion statistics are computed. The final motion representation is obtained by concatenating these motion statistics. The motion representation is then used for similarity measures between two videos of observed identities.

Chapter 3

Technical Background

In this chapter, we provide an overview of the theory, which is crucial for understanding the person Re-Id approaches described in the following chapters.

We start with the description of the terms Deep learning, Artificial Neuron, and describe a general Artificial Neural Network. Later, we continue with the Convolutional Neural Networks, Recurrent Neural Networks, Long-Short Term Neural Networks and ConvLSTM Neural Networks.

3.1 Deep learning

Deep learning is part of Machine Learning, which is based on Artificial Neural Networks [4]. The idea of the Artificial Neural Networks (ANN) first emerged in the 1940s after McCulloch and Pitts introduced simplified neurons in 1943 [32]. The ANNs are inspired by the biological neural networks that are present in the brain. The reason why the ANNs are gaining much attention in the recent years is that they have been shown to outperform previous state-of-the-art techniques in several tasks and together with the increasing amount of available data in fields like computer vision, they are believed to have a great potential in the future.

3.1.1 Artificial Neuron

The Artificial Neuron is the elementary unit of every ANN. It represents a function f, which processes an input $\mathbf{x} = [x_1, x_2, ..., x_n]$ such that

$$f(\mathbf{x}) = \varphi(F(\mathbf{x})) = \varphi(\sum_{i=1}^{n} w_i x_i + b). \tag{3.1}$$

In the above equation, the function φ is a so-called activation function, which decides whether a neuron should be activated and what is its relevance. The main purpose of the activation function is to introduce non-linearity into the output of a neuron. This is important as most real-world problems are not linear and the neural networks need to be able to learn these non-linear representations. The scalars $w_1, ..., w_n$ are learnable weights and scalar b is a learnable bias.

The neurons in a neural network are connected by connections and the learnable weights represent the significance of this connections. Each input to a neuron consists of outputs of previous neurons that have a connection with the current neuron. Figure 3.1 shows a diagram of a single neuron.

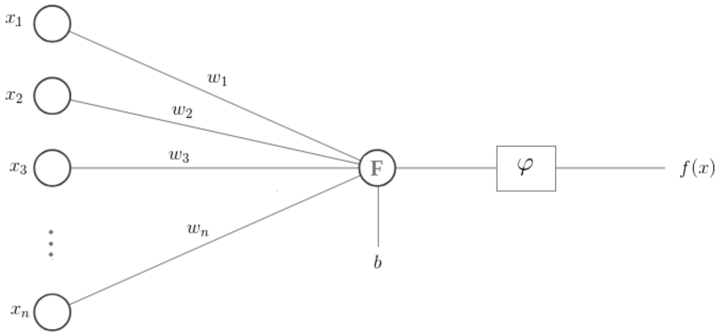


Figure 3.1: Neuron

3.1.2 General Artificial Neural Networks

Every ANN consists of an input layer, one or more hidden layers, and an output layer. Every layer contains multiple neurons, and they have connections to neurons from adjacent layers. In case all neurons from one layer are connected to all neurons from the previous layer, we talk about a dense (or fully connected) layer. An example of a neural network with two hidden layers can be seen in Figure 3.2.

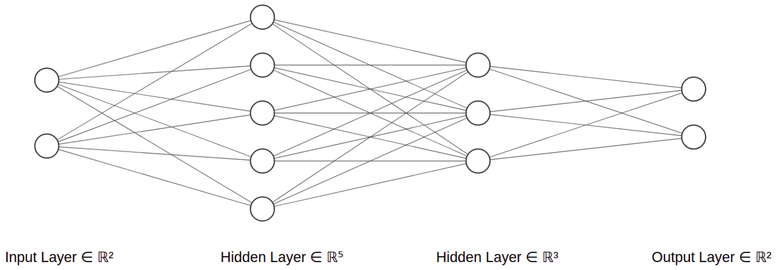

Input Layer ∈ $\mathbb{R}^2$
Hidden Layer ∈ $\mathbb{R}^5$
Hidden Layer ∈ $\mathbb{R}^3$
Output Layer ∈ $\mathbb{R}^2$

Figure 3.2: Artificial Neural Network diagram

3.1.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a particular type of ANNs that were originally designed for the usage in image processing applications. The characteristics of CNNs is the usage of convolution in at least one of its layers.

The reason why CNNs are so widely used in computer vision is that even small-sized images contain a tremendous amount of information. A grey-scale 620×480 image contains 297 600 pixels. Assuming that every pixel intensity of this image is an input to a fully-connected network, each neuron of this network requires 297 600 weights. Hence, the number of free parameters in the network becomes extremely large as the image size increases. This leads to over-fitting and slow performance. In CNNs, the total number of free parameters is reduced using the convolutional layers.

Another reason for the usage of CNNs is its translation invariance. In many pattern detection tasks, the same pattern can be found in different places in the image, and it would be inefficient to train neurons to recognize the same pattern on different positions independently.

A CNN usually consists of three types of layers: convolutional layers, pooling layers, and fully connected layers.

Convolutional Layers

In the convolutional layers, every neuron represents a convolution of a filter (called kernel) and a small patch in the image. The neurons are applied along the width and height dimensions of the image. The kernels have the same depth as the images, but smaller width and height. The output width and height of each layer depend on the kernel size, number of strides (number of pixels that the kernel is shifted between each computation) and the number of zero-padding pixels around the input image.

The main trick of the convolutional layers is the fact that the trainable weights (filters) are shared among a number of neurons in one layer. Hence, the number of free parameters in the convolutional layer is significantly lower than in a fully connected layer. The output of a convolution of the input with one filter (all neurons sharing the same weights) is called a feature map. Each filter in one layer generates one feature map. Concatenating the feature maps for all filters used in the layer crates the depth of the output of the layer. Every entry in the feature map can be interpreted as the output of one neuron.

Figure 3.3 displays a mapping of a convolutional layer. Every patch from the input on the left-hand side (one example is highlighted by the red block) is convolved with all filters of the layer. The convolutions of all patches from the input with one filter produce one feature map (the red plain on the right highlights one of them). The whole output consists of all feature maps, where the number of feature maps corresponds to the number of filters in this layer.

Every feature map detects some feature in the input. For example, in image processing, one feature map can detect horizontal edges, another one vertical edges, etc.

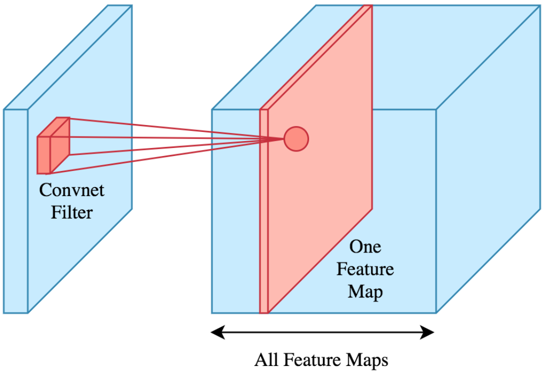
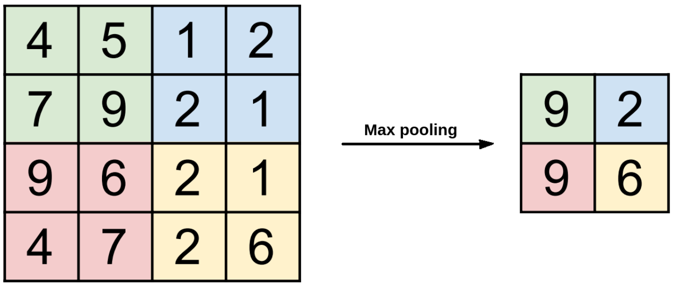


Figure 3.3: Convolutional layer [2]

Pooling Layers

The pooling layers progressively reduce the size of the data passed through them, and hence, they reduce the number of parameters of the network. There are several non-linear functions that can be used for pooling. The most common one is max pooling, where the input is partitioned into rectangles from which the max value is selected (see Figure 3.4).



4	5	1	2
7	9	2	1
9	6	2	1
4	7	2	6

9	2
9	6

Figure 3.4: Max pooling

Fully Connected Layers

In the CNNs, the fully connected layers are usually present at the end of the network. They are fully connected to all activations in the previous layers and perform the high-level reasoning of the network such as classification based on the features extracted by the preceding layers etc.

3.1.4 Recurrent Neural Networks

The traditional feed-forward neural networks are a powerful tool with many applications. However, one of their shortcomings is that when applied on sequential data, they fail to capture the sequential structure as they are not able to reason about the output based on the previously seen outputs. This is where the Recurrent Neural Networks (RNNs) come into play. They are networks with loops in them, allowing the output to be influenced not only by the weights that are learned during the training process but also by a state vector representing the context based on previously seen inputs and outputs.

The schema of a Recurrent Neural Network is illustrated in Figure 3.5. Here, a chunk of a neural network, A, receives the input x_t and produces the output h_t. The loop allows information to be passed from one step of the network to the next one. We can also interpret the RNN as a chain of multiple identical network models, each passing a message to the following one (see Figure 3.6).

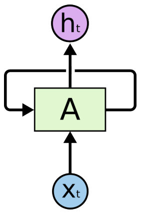


Figure 3.5: Recurrent neural network [5]

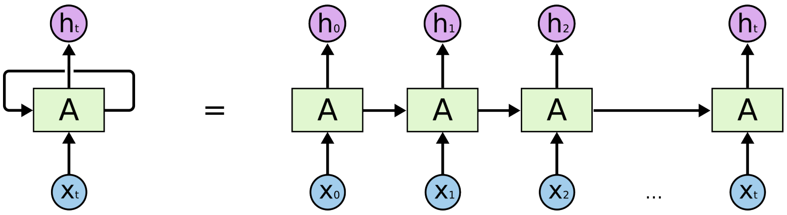


Figure 3.6: Unrolled recurrent neural network [5]

The problem of Long-Term Dependencies

In some cases, it is enough for the network to only remember the last few outputs from a sequence for good reasoning about the newly seen input. However, there are cases where the network needs to look far in the history to be able to classify the output for a given input correctly. In the case of the standard RNNs, the repeating modules have a very simple structure. Usually, they consist of one *tanh* layer (see Figure 3.7). In theory, despite their simple structure, RNNs are perfectly able to learn this so-called Long-Term Dependencies. However, in practice, they are not performing well in such situations (see [8] for more information). To overcome this issue, in 1997, Long Short-Term Memory Neural Networks were introduced [43].

3.1.5 Long Short-Term Memory Neural Networks

The Long Short-Term Memory Neural Networks (LTSMs) are a special type of RNNs, which are designed to avoid the Long-Term Dependency problem.

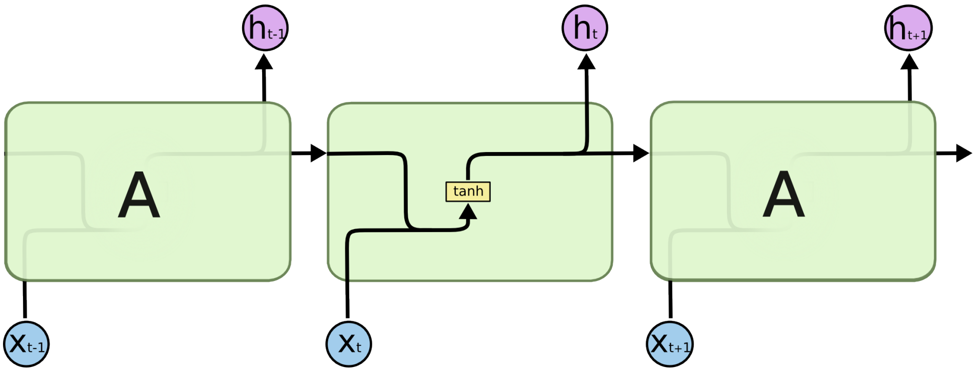


Figure 3.7: Repeating module in standard RNNs [5]

Unlike the standard RNNs, LSTMs are able to remember information for long time periods.

The key difference between the standard RNNs and LSTMs is the structure of the repeating network modules. Instead of having one single neural network layer, the schema of the LSTM repeating module is more complicated (see Figure 3.8).

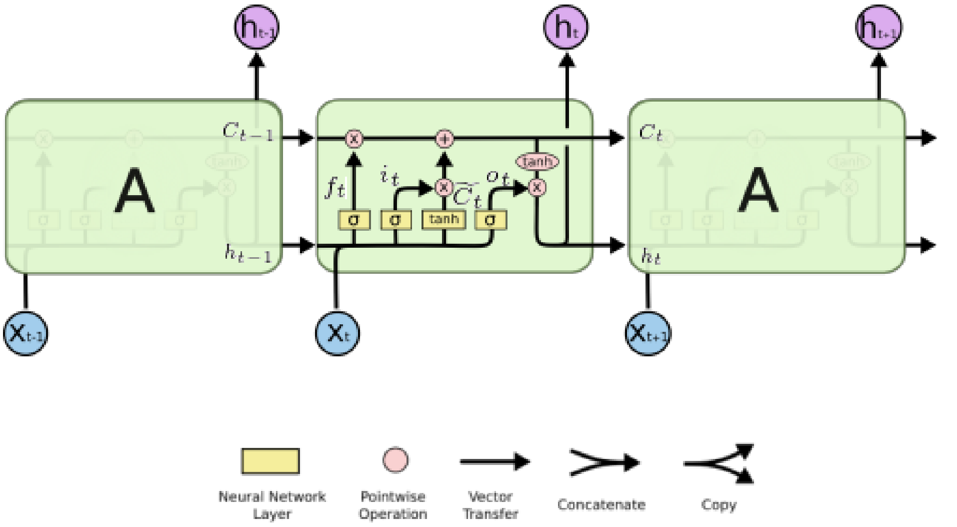


Figure 3.8: Repeating module in standard LSTMs [5]

It consists of two parts. The first part is the cell state (the horizontal line passing through the pointwise addition and pointwise multiplication nodes), which can be thought of as a transporter running down the whole chain of the network modules.

The second part is composed of three so called gates and one *tanh* layer. The gates consist of a sigmoid layer and a pointwise multiplication operation and output a number between zero and one, which specifies how much of the component should be remembered. The *tanh* layer, on the other hand, creates new candidate values, which can be to some extent (depending on the gates) added to the cell state.

The formulation of the LSTM module can be summarized by following equations.

$$f_t = \sigma(W_{xf} \cdot x_t + W_{hf} \cdot h_{t-1} + b_f) \tag{3.2}$$
$$i_t = \sigma(W_{xi} \cdot x_t + W_{hi} \cdot h_{t-1} + b_i) \tag{3.3}$$
$$\widetilde{C}_t = tanh(W_{xC} \cdot x_t + W_{hC} \cdot h_{t-1} + b_C) \tag{3.4}$$
$$C_t = f_t \circ C_{t-1} + i_t \circ \widetilde{C_t} \tag{3.5}$$
$$o_t = \sigma(W_{xo} \cdot x_t + W_{ho} \cdot h_{t-1} + b_o) \tag{3.6}$$
$$h_t = o_t \circ tanh(C_t) \tag{3.7}$$

where $\circ$ represents a pointwise multiplication, $\cdot$ a scalar multiplication and $+$ represents a vector addition. For detail information about the LSTMs, see [5].

3.1.6 ConvLSTM

We have already explained the Convolutional Neural Networks, which are networks, that are able to learn the spatial dependencies in images, and LSTM networks, that cover the temporal dependencies. However, one might be interested in covering both the spatial and the temporal dimension (e.g., when processing video sequences). For this case, the Convolutional LSTMs (ConvLSTMs) [50] were designed.

The ConvLSTMs follow the same structure as the LSTMs, but instead of a matrix multiplication, they use convolution operation inside of the LSTM module. Hence, the input and output data of the ConvLSTM are three-dimensional vectors, unlike in the standard LSTM module where the input and output data are one-dimensional. This way, ConvLSTMs are able to capture the underlying spatial features.

The formulation of the ConvLSTM module can be summarized by equations (3.8) to (3.13).

$$f_t = \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + b_f) \tag{3.8}$$
$$i_t = \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + b_i) \tag{3.9}$$
$$\widetilde{C}_t = tanh(W_{xC} * x_t + W_{hC} * h_{t-1} + b_C) \tag{3.10}$$
$$C_t = f_t \circ C_{t-1} + i_t \circ \widetilde{C_t} \tag{3.11}$$
$$o_t = \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + b_o) \tag{3.12}$$
$$h_t = o_t \circ tanh(C_t) \tag{3.13}$$

where $*$ represents a convolution operation. For more information about the convLSTMs, see [50].

Chapter 4

Gait as a biometrics

In this chapter, we provide an overview of body measurements, that can be used for a person-identification. In particular, we give intuition of why gait seems to be a promising feature that can be used for long-term Re-Id and describe the families of methods for gait recognition. This chapter was inspired by [20].

4.1 Biometrics

The term Biometric refers to a metric related to some human characteristics (biometric identifier). The biometric identifiers can be divided into two groups:

- Physiological identifiers (e.g., fingerprints, palm veins, DNA, iris...)
- Behavioral identifiers (e.g., voice, gait, typing rhythm...)

There are many human traits, that can be used for the biometric identification. Jain et al. [25] described seven factors that should be considered when selecting a particular biometric to be used in a specific situation:

- *Universality* - every person possess this characteristic
- *Uniqueness* - the characteristic is unique for each person
- *Permanence* - the trait does not vary over time
- *Measurability* - the measurements are easy to obtain
- *Performance* - relates to the accuracy, speed, and robustness of the technology used
- *Acceptability* - the technique is accepted well by the individuals of the relevant population such that they are willing to have their biometric trait captured
- *Circumvention* - the trait cannot be easily imitated

Currently, the most popular biometric-based person identification methods are based on iris recognition, face recognition, and fingerprints recognition. These methods are, in general, very accurate, fast, and robust. However, their main disadvantage is that they either require physical contact (fingerprint scanning) or at least a cooperation of the subject (photo shooting from a short distance and a suitable angle for iris or face recognition).

Gait, as a human trait, does not have these constraints and can easily be observed even from a distance. The studies in psychophysics [15] show that humans can recognize people by the way they walk, which indicates the presence of identity information in their gait. According to a medical study from the year 1967 [39], there are 24 different components of a human gait, and if all of them are considered, gait is unique for each person. For these reasons, the usage of gait to distinguish people is nowadays an attractive topic among researchers.

4.2 Approaches to Gait Recognition

Based on how the information about the individual's gait is obtained, gait recognition methods can be divided into three categories.

4.2.1 Wearable Sensor-Based Gait Recognition

In the Wearable Sensors-Based Gait Recognition methods, the gait information is collected using body-worn motion recording sensors. The sensors can be fastened at different locations on the human body and can hence measure various metrics such as speed, movement frequency, the maximal and minimal distance between specific body part, etc. These methods can be, depending on the sensors, very precise in capturing the human body movements. The apparent disadvantage, on the other hand, is the need of the sensors to be fastened onto bodies of the observed identities.

4.2.2 Floor Sensor-Based Gait Recognition

In the case of the Floor Sensor-Based Gait Recognition methods, a set of sensors is installed on the floor. These sensors enable to measure gait features such as stride length, stride cadence, or time on toe vs. time on heel ratio when a person walks on the floor. The main advantage of this method is that it does not need any cooperation from the side of the observed identities. The sensors can be installed in front of the doors of buildings and provide information about the gait of people who want to enter. On the other hand, sensors located on the floor can hardly monitor other body parts than legs and feet, and the person identification is then carried out on incomplete data.

4.2.3 Machine Vision-Based gait recognition

In the Machine Vision-Based Gait Recognition methods, the gait is captured by a video-camera from a distance. Various processing techniques are then applied to the videos to extract gait features for recognition purpose. The main advantage of this method is that the only necessary device that needs to be installed to collect the information about the target people is a surveillance camera, which is nowadays common equipment in many public places. The disadvantage of this method is its computational complexity as the processing of video sequences is, in general, significantly more computationally demanding than processing signals from sensors from the previous two methods. However, thanks to the increasing computational power available in devices, this obstacle is becoming less and less significant, and the machine vision-based gait recognition methods are believed to have significant potential in the field of the person recognition.

4.3 Challenges

Several factors may have a negative influence on the accuracy of gait recognition methods. We can divide them into two groups.

- *Internal factors.* These factors change the natural gait due to some physiological changes in the body such as pregnancy, sickness, injury, gaining or losing weight, drunkenness, aging and so on.

- *External factors.* These factors either impose challenges to the recognition algorithm (e.g., insufficient lighting conditions, varying viewing angles or indoor vs. outdoor environments) or temporary change the natural gait such as walking surface conditions (grass vs. concrete, dry vs. slippery floor, etc.), type of shoes (mountain shoes vs. high-heel shoes), objects carrying (e.g., backpack, suitcase, etc.) and so on.

Due to all these factors that need to be considered and managed in order to build a robust gait recognition system, for the time being, gait cannot be considered as a replacement for traditional person identification mechanisms like fingerprints or iris-based person identification. However, there are known cases where gait analysis was used as evidence in criminal investigations. One example is the investigation of an armed robbery in 2000 in the UK [7].

Chapter 5

Datasets

In this chapter, we provide an overview of the publicly available datasets for the person Re-Id. We then describe in details the MARS dataset, which we used in our experiments, together with the cleansing process that we applied to select the most suitable video sequences from all videos of the MARS dataset.

5.1 Overview

There are several publicly available datasets for the person Re-Id. A comprehensive overview can be found in [22]. One of the first was the ViPeR dataset [19] released in 2007. It contains 1264 images of 632 identities. Since then, a few more small-scale datasets appeared. The first dataset large enough for deep learning was the CUHK03 [34] introduced in 2014. It includes 13164 images of 1360 persons. One of the most popular person Re-Id datasets is the Market-1501 dataset [35] from the year 2015. It consists of 12936 images of 751 persons. This dataset was later extended into a MARS dataset [36], which is the first large scale video-based person Re-Id dataset [22]. This dataset was used for our experiments, and we describe it later in this chapter in more details. To my knowledge, the biggest Re-Id dataset with respect to the number of identities to date is the MSMT17 dataset [37]. It consists of 4101 identities and 126441 images. This dataset, however, does not contain tracking sequences and so it could not have been used for our purposes.

5.2 MARS Dataset

For the purpose of this thesis, we have used the MARS dataset [36], which is the biggest video-based person Re-Id dataset to date [22]. It contains 1261 identities and 1191003 video frames from around 20715 videos. All the recorded videos come from six near-synchronized cameras in the campus of Tsinghua. An example of a MARS video sequence consisting of seven image frames is displayed in Figure 5.1.

Even though MARS dataset is a suitable dataset to be used as a train-

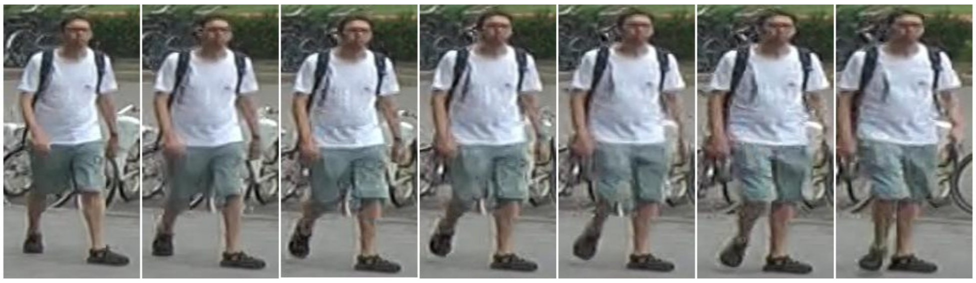


Figure 5.1: MARS data set example

ing dataset for the Person Re-Identification from video, there are some issues which have to be solved before using it for the model training.

5.2.1 Issues of the MARS data set in Re-Id

The main problem of the raw MARS dataset is the so called Class Imbalance Problem [27], which in case of the MARS data set means that the number of video sequences per identity varies significantly, ranging from 1 to 271 videos. Hence, when training a classification model, the identities for which there are many video sequences in the training dataset would be prioritized by the model which would harm its generalization ability.

Another issue is the length of the videos, which ranges from very short sequences consisting of less than five image frames up to very long sequences depicting dozens of steps of a walking person. If we want to build a model that should re-identify an identity based on the way they walk, the very short videos will not provide enough information for the model to be able to recognize them.

The third problem that needs to be mentioned when talking about the MARS dataset is the quality of the videos, which is also varying throughout the dataset. The most frequent defect of the videos is the partial occlusion, which limits the possibility to describe the movement of the whole bodies.

5.2.2 Dataset cleansing

In order to solve the previously mentioned issues, we apply the following cleansing process to the MARS dataset. We cut the original videos into sequences of 20 images, which roughly corresponds to a video of two steps of a walking person. The reason why we decided for the sequences of 20 images is that even though we would extract more information about the identities from longer sequences, we would loose a lot of identities as for many of them, there is not enough data to extract sufficient amount of longer video sequences. From these short sequences, we select up to 10 sequences for each identity. The selection is based on a parameter that denotes the quality of the sequence with respect to visibility of the moving identity. This step will be further explained in Chapter 6. Out of these ten sequences for one identity, at most 4 come from the same video as such sequences are very similar and would not bring any additional value for the model training. Those identities where there are less than 4 such video cuts are dropped from the data set. This way, we obtain a data set consisting of 1209 identities. The distribution of number of identities for number of sequences in the cleansed data set is displayed in diagram 5.2.

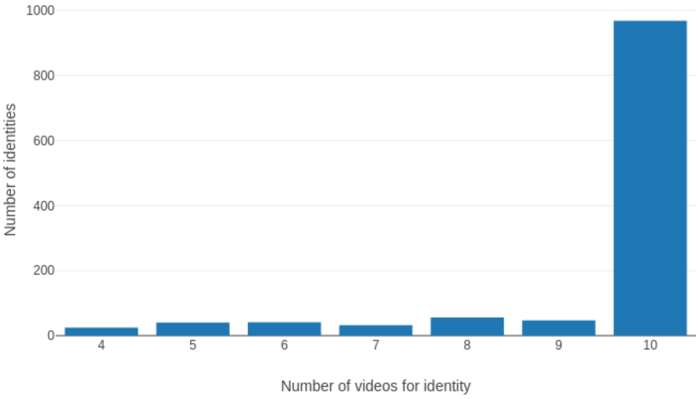


Figure 5.2: Distribution of number of identities for number of videos

Chapter 6

Proposed solutions

In this thesis, we use deep learning-based methods to solve the problem of Long Term Person Re-Identification. We propose three approaches: Edge-Based method, Skeleton-Based method, and Combined method. In all of these methods, the first step is the preprocessing of the video sequences. As each video consists of a sequence of images, we describe the preprocessing on individual image frames. The outputs of the preprocessing phase are then fed into a neural network. For the model implementation, we use the Keras library [14].

6.1 Edge-Based Method

In the Edge-Based Method, the preprocessing phase consists of a transformation of the original video frames into gray-scale images depicting the edges from the original images. We use this approach to remove information about a person's clothing, which typically changes in the long-term scope. The model is therefore forced to use information about the movement and body shape of the persons to perform the Re-Id. The transformed videos are then fed into a neural network described in Section 6.2.1.

6.2 Edge detection

In the edge detection problem, the task is to find edges and object boundaries in raw images. This problem is of a great importance to a variety of computer vision areas. Hence, a great number of approaches to edge detection have been developed, ranging from early developed methods using the Sobel operator [6] or Canny detector [18] to recently developed methods based on Convolutional Neural Networks such as DeepEdge [21] or CSCNN [28]. For the video frames transformation in the preprocessing phase of the Edge-Based method, we use the Holistically-Nested Edge Detection library (HED) available at [49], which provides an image-to-image method transforming the raw image into image of the objects boundaries.

The method is built upon a deep learning model that leverages fully convolutional neural networks and deeply-supervised nets [13], adopting a slightly modified VGGNet architecture [29]. An example of the HED image transformation applied on one image frame from the MARS data set can be seen in Figure 6.1. For more information about the HED library, see [49].

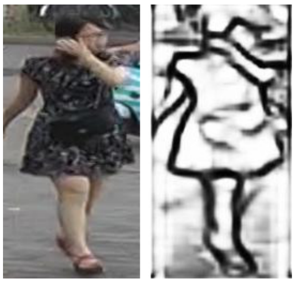


Figure 6.1: Edge extraction example: left-hand side - original image, right-hand side - image after edge extraction

6.2.1 Re-Id Neural Network description

The structure of the neural network used for the Person Re-Identification in the Edge-Based Method is illustrated in Figure 6.2.

The first part of this neural network consists of three stacked ConvLSTM layers (see Section 3.1.6), each of them followed by a batch normalization layer [1]. The input to the first ConvLSTM layer is a sequence of 20 images with the resolution of 64×64 pixels. The output of the last ConvLSTM layer is then fed into the second part of this network.

The second part of the network consists of an 2D Average Pooling layer, followed by a Flatten layer, a Dropout layer with a 50% dropout and two Dense layers with ReLu and Softmax activation functions.

The network uses categorical cross entropy [3] as its loss function.

6.2.2 Weak points

The main drawback of this method is the fact that when extracting the edges during the preprocessing phase, we keep the information about the shapes of the persons. However, the shape of the moving body changes with the clothes they are wearing. Hence, the same identities can look very different in two different videos.

The second problem of this method is that after the preprocessing phase, the contours of other people or objects remain in the images in case they were clearly visible in the original image. These undesired edges can be confusing for the model.

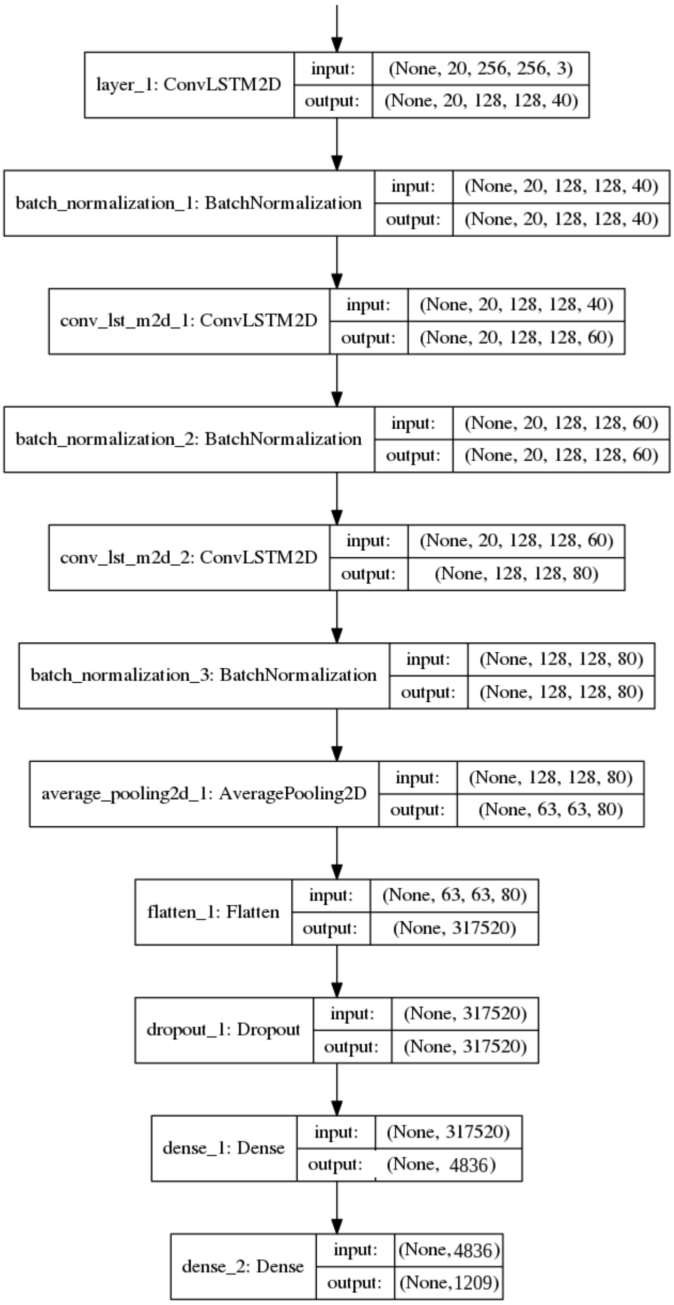

layer_1: ConvLSTM2D
input: (None, 20, 256, 256, 3)
output: (None, 20, 128, 128, 40)
batch_normalization_1: BatchNormalization
input: (None, 20, 128, 128, 40)
output: (None, 20, 128, 128, 40)
conv_lst_m2d_1: ConvLSTM2D
input: (None, 20, 128, 128, 40)
output: (None, 20, 128, 128, 60)
batch_normalization_2: BatchNormalization
input: (None, 20, 128, 128, 60)
output: (None, 20, 128, 128, 60)
conv_lst_m2d_2: ConvLSTM2D
input: (None, 20, 128, 128, 60)
output: (None, 128, 128, 80)
batch_normalization_3: BatchNormalization
input: (None, 128, 128, 80)
output: (None, 128, 128, 80)
average_pooling2d_1: AveragePooling2D
input: (None, 128, 128, 80)
output: (None, 63, 63, 80)
flatten_1: Flatten
input: (None, 63, 63, 80)
output: (None, 317520)
dropout_1: Dropout
input: (None, 317520)
output: (None, 317520)
dense_1: Dense
input: (None, 317520)
output: (None, 4836)
dense_2: Dense
input: (None,4836)
output: (None,1209)

Figure 6.2: Edge-Based model NN

6.3 Skeleton-Based Method

In the Skeleton-Based Method, the whole images are transformed into a structure that describes positions of seventeen joints describing the skeleton of the body. As the person walks, the changes of positions of their joints are representing their gait. The sequence of joints locations is the only input to the model. Hence, the model does not receive any information about the person's appearance and needs to rely solely on their movements.

6.3.1 Skeleton extraction

For the location and extraction of joints of the observed people in the videos in the Skeleton-Based Method, we utilize the the Openpifpaf library (available at [45]), which is currently the state-of-the art in the pose detection. This library is based on the work by Sven Kreiss at al. [44]. It provides a bottom-up method for multi-person 2D human pose estimation.

The whole main method of the library is based on a ResNet [30] network with two head networks: PIF (Part Intensity Field), which localizes the body parts, and PAF (Part Association Field) used to associate body parts with each other so that they form full human poses. The whole model is displayed in Figure 6.4. The input is an image of size (H, W) with three color channels. The neural network based encoder produces PIF and PAF fields. The decoder is then applied to convert the PIF and PAF fields into pose estimates containing 17 joints each. Each joint is represented by an x and y coordinate and a confidence score.

The output of the method is the image with marks on the positions of the 17 most significant joints of this person and lines connecting this joints and highlighting the whole skeleton. There are two modes in which this methods works: single-pose estimation, which selects the most reliable person with respect to their visibility in the image and highlights the skeleton for them only, and multi-pose estimation, which estimates the pose of all identities that are visible in the picture. For more information, see [44].

For our purpose, we use the single-pose estimation mode as we only are focused on one person in the image and we suppose they will always be the most visible one. We also need to slightly modify the outputs of the model so that we only obtain the skeletons without the original image. There are two possibilities to do so. The first one is to only extract the x and y coordinates of the 17 joints in the image which results in a 34 dimensional vector describing the whole skeleton. The second option is to draw the skeleton onto a plain white background without the original image. Figure 6.3 shows and example of a ten-image sequence of extracted skeletons. To keep some additional information, we draw the lines depicting hands blue, lines depicting the body green and those highlighting the legs red. Both can be easily obtained with some small modifications in the source code of the Openpifpaf library. The sequences of either the 34 dimensional vectors of joints locations or the sequences of the images displaying the skeletons are then used as the input to the corresponding neural network described in the following section.

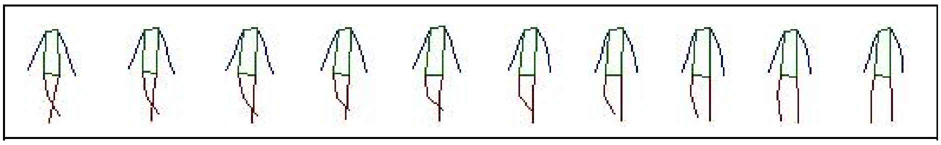


Figure 6.3: Sequence of extracted skeletons [44]

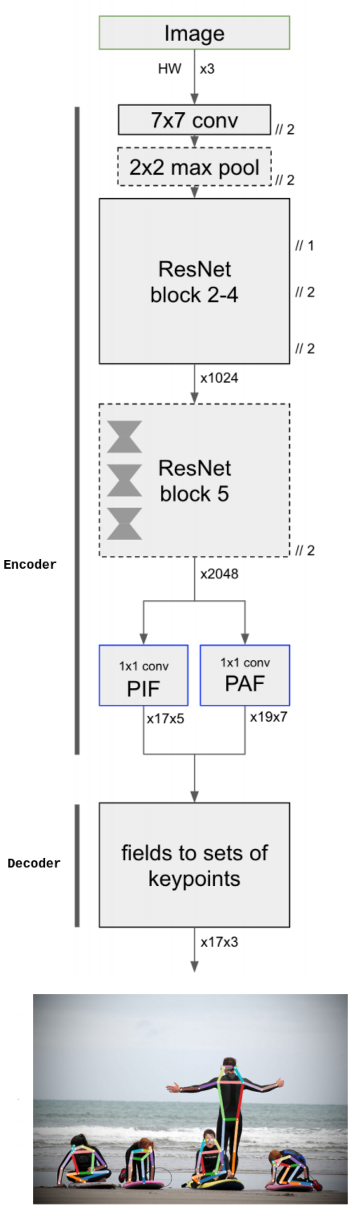


Figure 6.4: Openpifpaf model [44]

6.3.2 Re-Id Neural Network description

As the previous paragraph suggests, based on the type of the sequences of extracted skeletons, we need to use two different neural networks for the Skeleton-Based Person Re-Identification. In the case where the skeletons extracted from the images are drawn on a plain white background, we use the same neural network as in the Edge-Based Method (see Figure 6.2) as the input types are identical in both of the methods. For the case where the input sequences consist of 34 dimensional vectors describing the joints locations, we need to use a different network as the input sequences do not consist of image frames any more. The structure of this network can be seen in Figure 6.5. It can be again separated into two parts.

The first part consists of three stacked LSTM layers (see Section 3.1.5), each of which is followed by a batch normalization layer. The input to the first LSTM layer is a sequence of 20 34-dimensional vectors representing the x and y coordinates of the 17 most significant joints of the observed body (see previous section). The output of the last LSTM layer is then fed into the second part of the network.

The second part of the network consists of two Dense layers. The first one has the ReLu activation function, and the second one is the classifying layer and uses the Softmax activation function.

The network uses categorical cross entropy [3] as its loss function.

6.4 Combined method

In the third method, we combine both previous approaches. First, we transform the images into edges. In the second step, we draw the skeletons onto the edges. Same as in the Skeleton-based method, we draw the lines depicting hands blue, lines depicting the body green and those highlighting the legs red. The model is then provided with sequences of images depicting the edges together with the skeletons. This way, the model itself decides about the relevance of this two information and can combine them accordingly.

An example of a seven-frame sequence with edges and skeleton drawn as described is illustrated in Figure 6.6.

6.4.1 Re-Id Neural Network description

In the Combined Method, the inputs have the same type as the inputs in the Edge-Based Method and so we again use the same neural network for the person re-identification (see Figure 6.2).

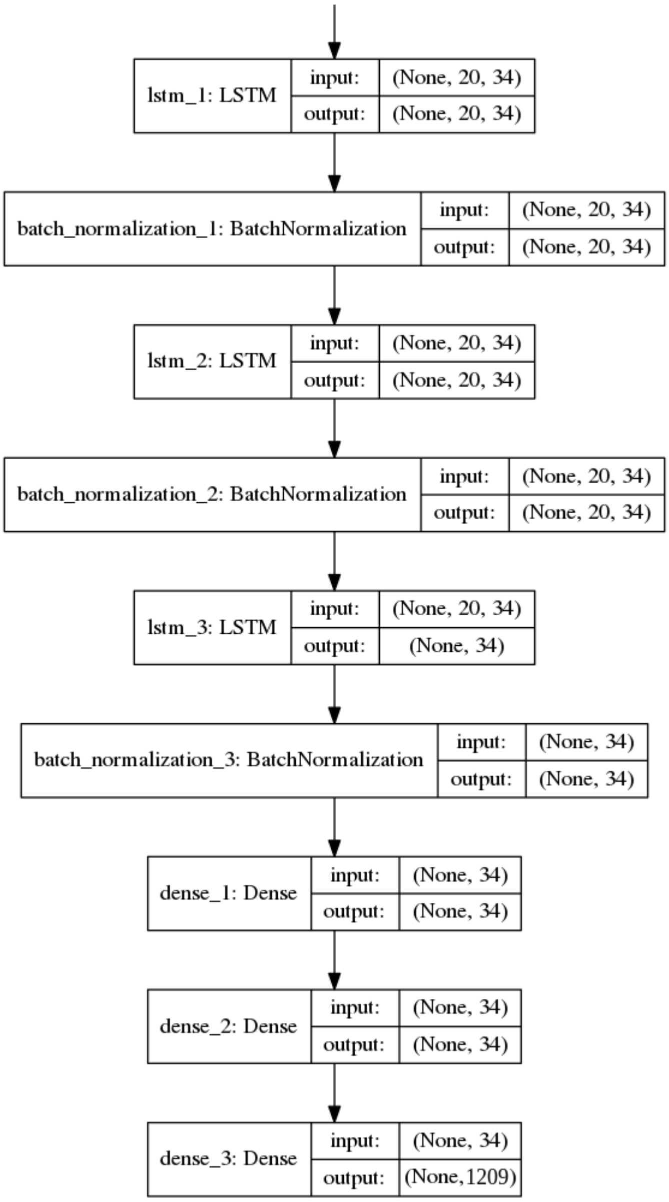


Figure 6.5: Skeleton-Based model NN

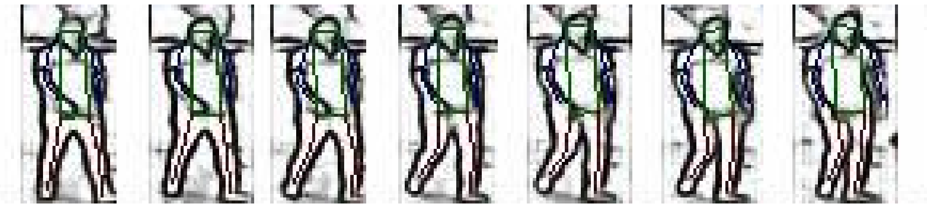


Figure 6.6: Edges with skeleton example

Chapter 7

Experiments evaluation

In this chapter, we present the experimental results of approaches for person re-identification described in previous chapter, together with the evaluation metric and train and test dataset split used.

7.1 Train and test data split

The process of obtaining the cleansed dataset used for our experiments was described in Section 5.2.2. After the cleansing process, we get a dataset consisting of 1209 identities. For each identity, there are between four and ten video sequences consisting of 20 image frames in our cleansed dataset. Additionally, at most four of these sequences come from the same video, as such sequences are too similar to bring any new information to the model if kept in a higher number.

This fact also needs to be taken into account when splitting the data into train and test sets. The sequences captured by one camera are never split between training and test dataset. Hence test data contain video sequences from angles and distances never previously seen by the model. Although this complicates the task, it more closely resembles the setting such model would face when deployed to real-world scenarios. To ensure this, we aggregate all the video sequences for all identities into groups where one group contains sequences obtained from one camera. These groups are then randomly split into train and test sets. In our experiments, we use 90% of the groups as the train data and 10% groups remain as the test data.

Due to the computation time and limited availability of the server used for training and evaluation, we could not use cross-validation to evaluate the models. To make the results of individual runs of all three algorithms described in Chapter 6 comparable, we used a fixed train and test data split.

7.2 Evaluation metrics

For the evaluation of our experiments, we use the standard model evaluation metric which is accuracy. The accuracy of the model measures the model performance and is defined as

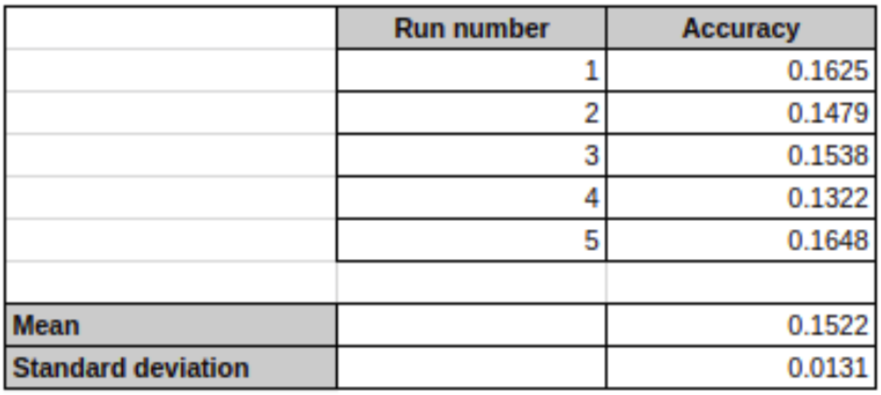
$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}. \tag{7.1}$$

7.3 Results

The following results were obtained from five runs of each of the algorithms described in Chapter 6. For the training, we used a server with 256GB RAM, 500GB SSD and a NVIDIA GTX 1080Ti GPU. Each training took approximately 40 hours.

7.3.1 Edge-Based Method

Table 7.1 provides the results of the five independent runs of the Edge-Based Method. We can see that on average, the trained models correctly predict the identity for a sequence of images in approximately 15.2% of the cases with the standard deviation of approximately 1.3%.



	Run number	Accuracy
	1	0.1625
	2	0.1479
	3	0.1538
	4	0.1322
	5	0.1648
Mean		0.1522
Standard deviation		0.0131

Table 7.1: Results of five runs of the Edge-Based algorithm

Figure 7.1 shows the Validation Accuracy learning curve for the Edge-Based Method. The x-axis represents the number of iterations of the algorithm and the y-axis denotes the accuracy.

Figure 7.2 shows an example of all miss-classifications for one sequence. For simplicity, we only display one image frame for each sequence. On the left-hand side, we see an image representant of correct class. The images on the right-hand side are then all representants of sequences wrongly classified

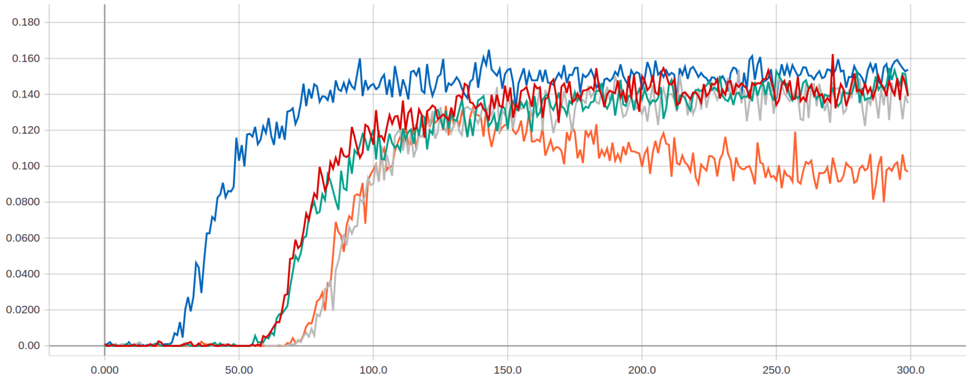


Figure 7.1: Validation Accuracy learning curve

as being of the left identity. We can see that the images share some features. In all the cases, the person in the image is wearing a skirt, which also implies that all of them are most probably women. We can also say that they all are rather skinny. Four out of five identities on the right-hand side have a very similar long hair as the identity on the left. This indicates that even though we tried to eliminate the visual features in the training data, the model still utilizes those that remain in the preprocessed videos. This motivates us to try an approach which completely hides such visual features and focuses only on skeletons of the observed identities.

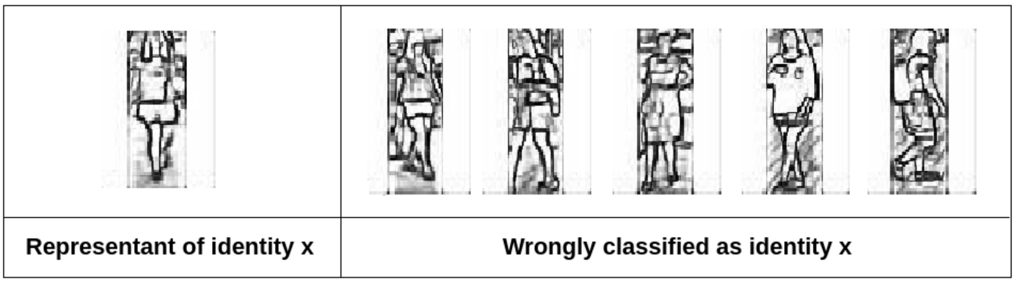


Figure 7.2: Edge-Based method miss-classifications example

7.3.2 Skeleton-Based Method

We tried two different approaches for skeleton based method (see Section 6.3), and so we should present two different results. However, neither of these two approaches provided any better results on the validation data set then random classification.

The main reason for these unsatisfactory results is the poor quality of the dataset, where the resolution of the videos is low, and the observed identities are often occluded. Also, the Openpifpaf library for skeleton extraction 6.3.1 is not perfect, and hence, it often did not succeed in the extraction of the whole skeleton. If we look at the results of the skeleton the extraction method, we can see that often we did not manage to extract the whole skeleton with all the body parts as some of the joints were not found in the images. Figure 7.3 shows an example of two ten-frame sequences of extracted skeletons for one identity. The first sequence is an example of a good quality sequence where all skeleton parts are visible. The second sequence misses some body parts. We can presume that sequences of such incomplete skeletons had a significant negative impact on the ability of the model to learn the walking patterns for individual identities.

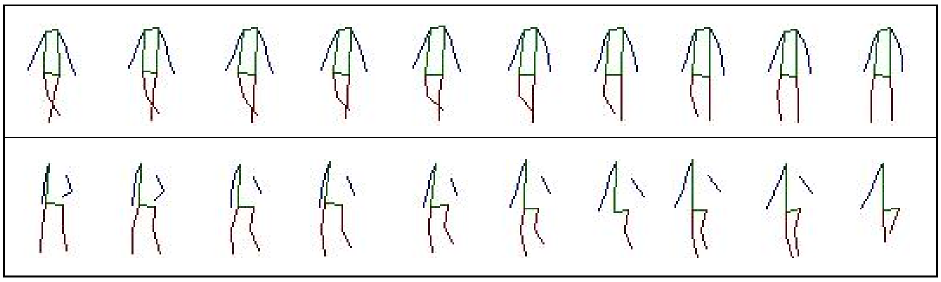


Figure 7.3: Bad quality skeleton sequence example

7.3.3 Combined Method

To verify whether the extracted skeletons carry any additional useful information for the task of person re-identification, we performed a set of experiments where we combine both previously mentioned approaches. Table 7.2 provides the results of the five independent runs of the Combined algorithm. We can see that on average, the trained models correctly predict the identity for a sequence of images in approximately 15.61 of the cases with the standard deviation of approximately 0.5%. Figure 7.4 shows the Accuracy learning curve for the Combined method. The x-axis represents the number of iterations of the algorithm and the y-axis denotes the accuracy.

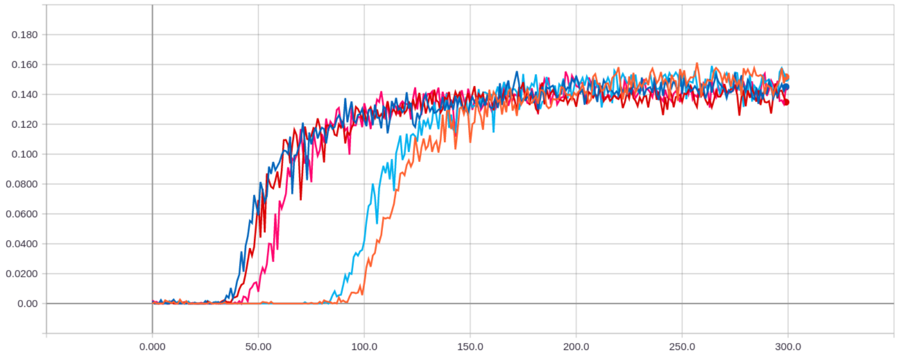
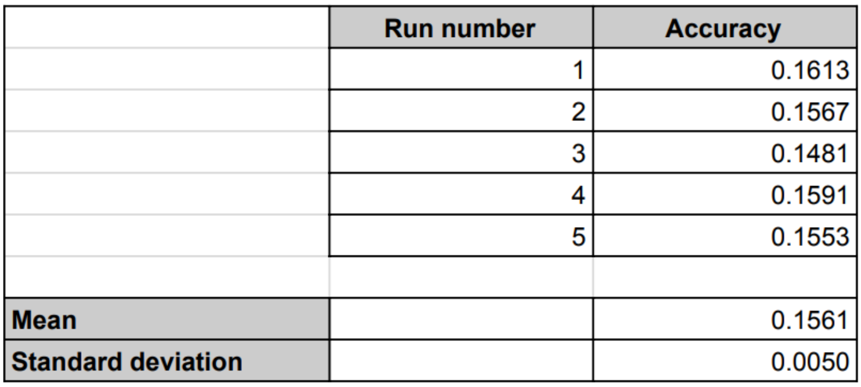


Figure 7.4: Validation Accuracy learning curves

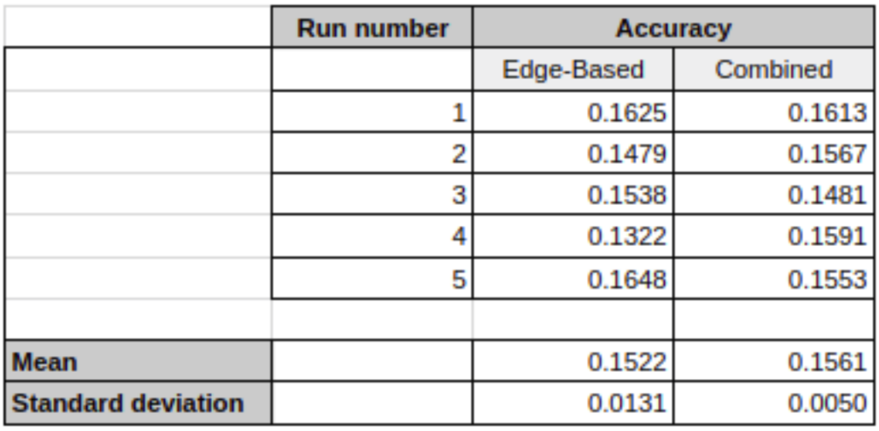


	Run number	Accuracy
	1	0.1613
	2	0.1567
	3	0.1481
	4	0.1591
	5	0.1553
Mean		0.1561
Standard deviation		0.0050

Table 7.2: Results of five runs of the Combined algorithm

Comparison with the Edges-based method

We will now compare the results of the Combined Method with the results of the Edge-Based Method. Table 7.3 shows the results of five runs of each of these two algorithms. Based on the mean of the accuracy and its standard deviation, we could say that the Combined Method performs slightly better than the Edge-Based Method. However, this little difference is caused by the fourth run of the Edge-Based algorithm, which shows significantly worse performance than the rest of its runs and can be considered as exceptional. The rest of the runs of both the algorithms result in a very similar validation accuracy. We can conclude that adding the information about the extracted skeletons of the identities did not help the model to identify the identities. This further strengthens the conclusion that current approach for skeleton extraction is insufficiently detailed and precise.



	Run number	Accuracy	
		Edge-Based	Combined
	1	0.1625	0.1613
	2	0.1479	0.1567
	3	0.1538	0.1481
	4	0.1322	0.1591
	5	0.1648	0.1553
Mean		0.1522	0.1561
Standard deviation		0.0131	0.0050

Table 7.3: Comparison of Edge-Based and Combined method

Chapter 8

Conclusion and Future Work

8.1 Summary

In this thesis, we studied the methods of person re-identification in videos. In particular, we focused on the Long-Term Person Re-Identification, where the methods used need to focus on non-visual features which do not change significantly with the time. The main contributions of this thesis are:

- **Overview of available methods for person re-identification**

  We provide an overview of methods for person re-id from the early beginnings, when person re-id was thought of as a part of the person tracking problem and matching of features extracted from individual image frames was used to perform the re-id, until nowadays, when advanced deep learning systems are applied on the whole video sequences in order to utilize the most possible information from the videos. Even though in this thesis we focus on the Long-Term Person Re-Id, the overview covers both long-term and short-term re-id approaches as these two umbrella terms are closely related, and approaches to short-term re-id can be a valuable source of inspiration for the long-term re-id methods.

- **Discussion of the gait as a biometrics**

  We examine the gait as a mean of person re-identification and compare it with other popular biometrics such as iris recognition, face recognition, or fingerprints recognition. We also describe three main approaches to gait recognition: Wearable Sensor-Based Gait Recognition, Floor Sensor-Based Gait Recognition, and Machine Vision-Based Gait Recognition.

- **A novel solution to Long-Term Person Re-Id**

  As part of this thesis, we implemented three methods that aim to solve the Long-Term Person Re-Identification Problem. This problem is especially challenging for two reasons. First of all, to my knowledge, the datasets available for the person Re-Id are not sufficient for training a model with a good performance. The video sequences from the available datasets are of poor quality with low resolution. Moreover, the identities in the videos are often partially occluded, and different video sequences of the same identity usually have similar visual features that are difficult to eliminate (same or similar clothes, hairstyle, etc.). On the other hand, video processing is always a costly operation requiring much memory and computational power, and training a model on a large dataset with videos of a high resolution would take long time even on a very powerful computer.

All the methods that we implemented use the MARS dataset for training and validation. In the first method, we extract the contours of identities in the video sequences in the dataset and train a classification ConvLSTM network. In the second method, we transform the videos into sequences of locations of 17 joints of the identities present in the videos and train a classification LSTM network on this sequences only. The third method combines both previous approaches.

The results of our three approaches are the following. The Edge-Based Method as well as the Combined method achieve the accuracy of approx. 15.5% on the test data set. The Skeleton-Based method does not achieve any higher accuracy than a random classification. The model was not able to extract from the joints locations any information, that would uniquely identify the observed identities.

This seemingly contradicts the assumption from Chapter 4, which says that humans can be uniquely identified by the way they walk. However, this assumption involves the consideration of 24 different components of a human gait, including parameters like foot angle, head movements, and other, that cannot be extracted from the locations of 17 joints describing the fundamental skeleton. Another reason for these unsatisfactory results is the pure quality of the videos from the MARS dataset. The identities are often occluded by other objects in the videos, and hence, some of the joints locations are missing in the sequences. Moreover, some of the video sequences contain identities that are riding a bicycle or motorcycle instead of walking, which makes it impossible for the model to extract any information about their gait. The last thing that should be mentioned is the length of the video sequences that we used for training and validation. To obtain enough training data, we cut the original videos into sequences of 20 images only. Longer sequences would probably provide more information. However, we would lose many identities as, for many of them, MARS data set does not contain a sufficient amount of long enough sequences.

8.2 Future Work

The research on long-term person re-identification is a relatively young area in computer vision. Many aspects are yet to be examined, and a large amount of work needs to be done to obtain a system that can reliably recognize and re-identify people.

The main area that needs to be explored is the gait recognition as it is believed to have great potential in the Long-Term Person Re-Id. The gait is easy to be recorded without the cooperation of the target person, it does not change significantly during the time, and it is believed to be unique for each person. With the increasing computational power, it is becoming easier to train models on a large number of high-quality videos and to use the trained models in real-time.

The apparent prerequisite for training a reliable model is the collection of an appropriate dataset, which is not yet publicly available. Such a dataset must contain videos for a high number of identities so that it can be used as a training dataset for deep learning methods. The videos must be of high quality such that the details of the body movements are clearly visible. The videos should be recorded from various viewpoints so that the model trained on such a dataset can be applied to videos recorded from general surveillance cameras. Finally, for each identity, there should be video records for various clothing or hairstyles such that the model is forced not to focus on the visual features.

One possibility to increase the performance of the gait recognition models is to combine the videos with records from some body-movement sensors, that can be built on the floor. They can provide additional information which cannot be obtained from the videos using the currently available method but is crucial for successful gait recognition.

Appendix A

Contents of the CD

```
/
├── personReId
│   ├── data
│   ├── src.................. Python code of the implemented methods
│   └── notebooks........... Jupyter notebooks used for data evaluation
└── thesis................................ LaTeXcodes of this thesis
    ├── figures
    └── chapters
```

Appendix B

Bibliography

[1] Batch normalization layer. https://keras.io/layers/normalization/.

[2] Convolutional layer. https://www.kdnuggets.com/2019/07/convolutional-neural-networks-python-tutorial-tensorflow-keras.html.

[3] Cross entropy. https://en.wikipedia.org/wiki/Cross_entropy.

[4] Deep learning. https://en.wikipedia.org/wiki/Deep_learning.

[5] Lstm networks. https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[6] Sobel operator. https://en.wikipedia.org/wiki/Sobel_operator.

[7] Rami Abboud, Richard Baker, and Julie Stebbins. Forensic gait analysis: A primer for courts, 2017.

[8] Anastasios Doulamis Eftychios Protopapadakis Athanasios Voulodimos, Nikolaos Doulamis. Learning long-term dependencies with gradient descent is difficult. *IEEE*, 1994.

[9] Anastasios Doulamis Eftychios Protopapadakis Athanasios Voulodimos, Nikolaos Doulamis. Deep learning for computer vision: A brief review. *IEEE*, 2017.

[10] Yu Su Bingpeng Ma and Frederic Jurie. Local descriptors encoded by fisher vectors for person re-identification. *IEEE*, 2012.

[11] Shaogang Gong Tao Xiang Bryan Prosser, Wei-Shi Zheng. Person re-identification by support vector ranking. 2010.

[12] L. Liu J.-G. Yu C. Gao, J. Wang and N. Sang. Temporally aligned pooling representation for video-based person re-identificatio. *IEEE*, 2016.

[13] Patrick Gallagher-Zhengyou Zhang Zhuowen Tu Chen-Yu Lee, Saining Xie. Deeply-supervised nets. *AISTATS*, 2014.

[14] François Chollet. keras. https://github.com/keras-team/keras, 2015.

[15] James E Cutting and Lynn T Kozlowski. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the psychonomic society*, 9(5):353–356, 1977.

[16] S. Liao S. Z. Li et al. D. Yi, Z. Lei. Deep metric learning for person re-identification. *ICPR*, 2014.

[17] N. Dalal and B. Triggs. Long-term person re-identification using true motion from videos peng. *IEEE*, 2005.

[18] Lijun Ding and Ardeshir Goshtasby. On the canny edge detector. *Pattern Recognition*, 34(3):721–725, 2001.

[19] Hai Tao Douglas Gray. Viewpoint invariant pedestrian recognition with an ensemble of localized features. *ECCV*, 2008.

[20] Davrondzhon Gafurov. A survey of biometric gait recognition: Approaches, security and challenges. In *Annual Norwegian computer science conference*, pages 19–21. Annual Norwegian Computer Science Conference Norway, 2007.

[21] Lorenzo Torresani Gedas Bertasius, Jianbo Shi. Deepedge: A multi-scale bifurcated deep network for top-down contour detection. *CVPR*, 2015.

[22] Mengran Gou. Re-id data set overview. https://github.com/NEU-Gou/awesome-reid-dataset, 2019.

[23] Cordelia Schmid Heng Wang. Action recognition with improved trajectories. *ICCV - IEEE*, 2013.

[24] D. Sagi I. Fogel. Gabor filters as texture discriminator. 1989.

[25] Pankanti Sharath Jain Anil, Bolle Ruud. *Biometrics*. 1999.

[26] Yann LeCun Eduard Säckinger Jane Bromley, Isabelle Guyon and Roopak Shah. Signature verification using a "siamese" time delay neural network. 1993.

[27] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

[28] Tyng-Luh Liu Jyh-Jing Hwang. Pixel-wise deep learning for contour detection. *ICLR*, 2015.

[29] Tyng-Luh Liu Jyh-Jing Hwang. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[30] S. Ren K. He, X. Zhang and J. Sun. Deep residual learning for image recognition. *IEEE*, 2016.

[31] Wei Zhang Rui Huang Kan Liu, Bingpeng Ma. A spatio-temporal appearance representation for video-based pedestrian re-identification. *IEEE*, 2015.

[32] Ben Kröse, Ben Krose, Patrick van der Smagt, and Patrick Smagt. An introduction to neural networks. 1993.

[33] A. Perina M. Farenzena L. Bazzani, M. Cristani and V. Murino. Multiple-shot person re-identification by hpe signature. *IEEE*, 2010.

[34] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014.

[35] Lu Tian Shengjin Wang Jingdong Wang Qi Tian Liang Zheng, Liyue Shen. Scalable person re-identification: A benchmark. *IEEE*, 2015.

[36] Yifan Sun Jingdong Wang Chi Su Shengjin Wang Qi Tian Liang Zheng, Zhi Bie. Mars: A video benchmark for large-scale person re-identification. *ECCV*, 2016.

[37] Wen Gao Qi Tian Longhui Wei, Shiliang Zhang. Person transfer gan to bridge domain gap for person re-identification. *IEEE*, 2018.

[38] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. *IEEE*, 2012.

[39] Murray MP. Gait as a total pattern ofmovement. *American Journal of Physical Medicine*, 1967.

[40] J. Martinez del Rincon N. McLaughlin and P. Miller. Recurrent convolutional network for video-based person re-identification. *IEEE*, 2016.

[41] Jingsong Xu Jian Zhang Peng Zhang, Qiang Wu. Long-term person re-identification using true motion from videos peng. *IEEE*, 2018.

[42] Cordelia Schmid. Constructing models for content-based image retrieval. *IEEE*, 2001.

[43] Juergen Schmidhuber Sepp Hochreiter. Long short-term memory. 1997.

[44] Alexandre Alahi Sven Kreiss, Lorenzo Bertoni. Pifpaf: Composite fields for human pose estimation. 2019.

[45] George Adaimi Sven Kreiss, Junedgar. Composite fields for human pose estimation. `https://github.com/vita-epfl/openpifpaf`, 2019.

[46] X. Zhu T. Wang, S. Gong and S. Wang. Person re-identification by video ranking. *European Conference on Computer Vision*, 2014.

[47] T. Xiao W. Li, R. Zhao and X. Wang. Deepreid: Deep filter pairing neural network for person re-identificatio. *IEEE*, 2014.

[48] Z. Zivkovic W. Zajdel and B. Krose. Keeping track of humans: Have i seen this person before? *IEEE*, 2005.

[49] Tu Zhuowen Xie, Saining. Holistically-nested edge detection. `https://github.com/s9xie/hed`, 2015.

[50] Hao Wang Dit-Yan Yeung Wai-kin Wong Wang-chun WOO Xingjian SHI, Zhourong Chen. Convolutional lstm network: A machine learning approach for precipitation nowcasting. 2015.

[51] Zhichao Song Chao Ma Yan Yan-Xiaokang Yang Yichao Yan, Bingbing Ni. Person re-identification via recurrent feature aggregation. *ECCV*, 2016.

[52] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.