DEGREE PROJECT IN TECHNOLOGY,
FIRST CYCLE, 15 CREDITS
STOCKHOLM, SWEDEN 2021

The ghost in the machine

Exploring the impact of noise in datasets used for graph-based action recognition

RICKY MOLEN

RUBEN REHN

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

The ghost in the machine

Exploring the impact of noise in datasets used for graph-based action recognition

RICKY MOLEN AND RUBEN REHN

Degree project, first cycle (15 hp)
Date: June 24, 2021
Supervisor: Hedvig Kjellström and Jörg Conradt
Examiner: Pawel Herman
School of Electrical Engineering and Computer Science
Swedish title: Skräp in, skräp ut - En utredning av bruskänslighet i graf-baserad rörelseigenkänning

Abstract

Human action recognition is the task of classifying human movement and actions from video data. To benchmark different algorithms within the action recognition field, a common benchmark dataset, called NTU-RGB+D is used. However, this dataset is not without its issues as some samples contain data that is mistakenly captured as a human. In the context of this thesis, these are defined as ghost bodies. This thesis explores to what extent the accuracy of a state-of-the-art directed graph neural net, DGNN, is affected if trained without ghost bodies. The results suggest that the accuracy increases by 1.79 percentage points when ghost bodies are excluded during testing with an unofficial implementation of the DGNN. However, the results of the original DGNN could not be fully replicated which undermines the strength of the results. Despite this, given the importance of the NTU dataset within action recognition, we suggest considering a new benchmark dataset that takes ghost bodies into account. While the results of the study are not generalizable, the measured difference in recognition accuracy still points to the the necessity of looking deeper into the phenomenon of ghost bodies within action recognition.

Sammanfattning

Mänsklig rörelseigenkänning (*en. human action recognition*) är forskningsområdet ägnat åt att känna igen mänskliga rörelser från videodata. För att kunna jämföra olika algoritmer inom området förekommer ofta ett standardiserat datasetet, NTU-RGB+D, som bland annat innehåller skelettrepresentationer av människor som utför rörelser. Trots datasetets vida användning inom rörelseigenkänning innehåller det vad som i denna uppsats benämns spökkroppar (*en. ghost bodies*). Dessa artefakter i datasetet är skelettrepresentationer som felaktigt klassats som att de tillhör en människokropp när de i själva verket utgör något annat icke-mänskligt objekt i videodatan. Experimentet som redogörs för i denna uppsats har ägnats åt att undersöka hur dessa spökkroppar påverkar rörelseigenkänningsprecisionen (*en. action recognition accuracy*) hos ett nutida riktad-graf-baserat neuralt nätverk (*en. directed graph neural network, DGNN*). Resultaten visar att igenkänningsprecisionen tycks öka med 1,79 procentenheter när grafnätverket tränas utan förekomster av spökkroppar. Resultaten bör dock tolkas med försiktighet då den igenkänningsprecision som rapporterats för grafnätverket i originalexperimentet inte kunde replikeras. Trots detta utgör NTU ett så pass viktigt dataset för forskning inom rörelseigenkänning, att vidare analys och förbättring av datasetet med avseende på spökkropparna är att rekommendera. Även om resultaten inte kan generaliseras bortom det grafnätverk som experimentet utfördes med, pekar ändå den uppmätta skillnaden i igenkänningsprecision på vikten av vidare analys vad gäller spökkroppars inverkan på moderna algoritmer inom rörelseigenkänning.

Contents

Chapter 1

Introduction

Computer science can be seen as the research area where one is aiming to understand the capabilities of computers. However, with the advent of artificial intelligence, much attention within computer science has been devoted to finding algorithms and methods for computers to understand humans rather than us understanding them. To interpret humans, a computer must be able to listen, comprehend human feelings and see humans interact with each other and the world around them.

There are many different research areas where efforts are made to replicate human abilities. Among them are the interpretation of human language (Natural language processing), understanding human emotion (Sentiment analysis) as well as giving computers the ability to see (Computer vision).

In the past decade, major breakthroughs have been made towards interpreting images and understanding visual content with the help of deep learning algorithms (see chapter 2). This has for example been used to classify and segment different parts and objects of an image. Computer vision has also been naturally extended to work with video data which have opened up the possibility for a large array of applications and research areas. One such research area is called video action recognition.

Video action recognition is the task of classifying, often human, physical actions in sequences of images [1, 2]. The applications for human action recognition are many including video surveillance and human-computer interaction tasks [3]. The problem of action recognition is still subject to active research and new approaches and improvements are proposed frequently [3, 4, 5, 6].

There are several different approaches for recognizing actions from video data. Recent work mainly revolves around so-called skeleton-based action recognition where joints and bone positions are extracted from images [3]. There have also been different propositions on how to best use skeleton data in order to recognize human actions. Firstly, sequence-based methods represent the human body as a sequence of joints. These are then trained using recurrent neural networks, RNNs [7]. Other approaches involve convolutional neural networks [8]. Lastly, graph neural networks, GNNs instead represent the body as joints and bones where joints and bones can be seen as nodes and edges in a graph respectively. Recent work within human action recognition leverage an extension of GNNs. These are called directed graph neural networks which has given remarkable results in contemporary studies. These different approaches are explained further in chapter 2.

The common factors among different action recognition algorithms are many. Perhaps the most important one revolves around how they are trained and evaluated. There are multiple datasets used for benchmarking these algorithms, and one of the most frequently used is the NTU-RGB+D dataset [7]. It contains, among other things, graph representations of human skeletons and introduces a standardized separation of training and testing data that can be used to compare the accuracy of different action recognition algorithms. However, this dataset, like any other, is not perfect. The skeleton representations in the dataset were created with the use of the Microsoft Kinect sensor. This sensor is used to identify human joints in video data. However, sometimes the sensor falsely detects human joints in other objects that are in fact not human. This results in unwanted noise which could potentially affect the accuracy of action recognition algorithms. An example of such noise can be seen in Figure 1.1, where a chair has falsely been identified as a human.

As the quality of datasets is such an important factor when training any machine learning algorithm, it is interesting to dive deeper into how action recognition algorithms might be affected by noise in the datasets they are trained on.

1.1 Definitions and fundamentals

In order to understand the problem and research question that will be presented in this thesis, some definitions and fundamental terms must first be understood. Those fundamentals will be explored within this section.

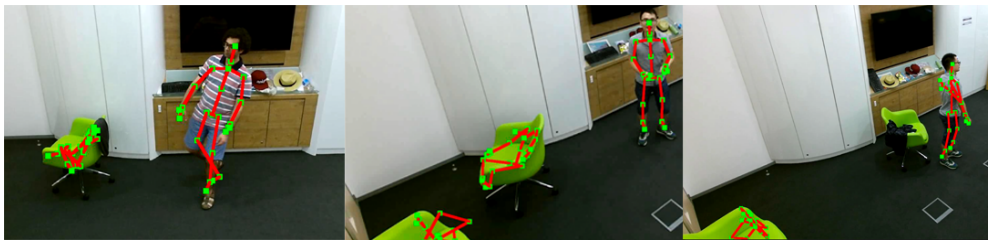

Figure 1.1: Example of samples with ghost bodies (see subsection 1.1.5) where one or more chairs are inserted as skeletons

1.1.1 Learning algorithms

In order to understand contemporary action recognition algorithms, it is important to understand the concept of machine learning. Tom M. Mitchell [9] coined the definition as *"A computer program is said to **learn** from experience E with respect to some class of task T and performance measure P. if its performance at tasks in T, as measured by P, improves with experience E."*. By experience E it is possible to imagine a high variety of interpretations but most commonly is some form of data. A task T, is harder to clarify considering a task can take many forms and sometimes non-intuitive tasks like memorizing data. But it is important to state that the T is not learning in itself. Rather, T refers to the problem that the algorithm is aiming to solve. Some common examples are classification, regression or machine translation. How well the algorithm performs, P, is the quantified measurement of how the algorithm is learning. Depending on the specific algorithm and what is interesting to measure, there are different ways of measuring P. For a classification task, it is normal to aim for a P measured on unseen data and this is managed by splitting the data into a test and training dataset where the model learns from the training data and evaluated on the test data.

1.1.2 Deep learning in action recognition

In the past couple of years action recognition using deep learning has increased dramatically. There are many definitions of deep learning, but in short, it can be described as a subset of machine learning where so-called artificial neural networks, ANNs, are employed to learn to infer from looking at data. Moreover, a deep neural net consists of several layers of artificial neurons that have proven remarkably accurate when it comes to learning from data in complex areas such as natural language processing and image recognition to name a few. [10].

Moreover, the majority of action recognition approaches today uses deep learning. However, these approaches differ in much of their implementation and how they learn even though they all employ some form of deep learning to recognize the action.

1.1.3 Graph-based action recognition

For human action recognition, many deep learning approaches use what is called skeleton-based action recognition. Skeleton-based methods involve extracting a representation of the human skeleton from video data, and using it as input for action recognition algorithms [3]. The skeleton-based action recognition up until this point are divided into three different categories. Each of these categories are explored in-depth in chapter 2.

1.1.4 The NTU-RGB+D dataset

The NTU-RGB+D [7] dataset is one of the most widely used datasets for action recognition and therefore commonly used to benchmark new algorithms against existing implementations. The dataset contains 60 action classes and 40 subjects where 11 of the classes involve two people. Some examples of classes are *take off a hat/cap, cheer up, phone call or point finger*. The dataset contains RGB data (red, green blue), depth data, skeleton data and infrared data. However, within the confines of this thesis only the skeleton data will be used. Skeleton data, often called joint data, simply consists of a set of 3D coordinates marking the location of a human joint in space. Thus, a set of these human joint coordinates can be seen as a representation of the entire human body. For samples in the NTU-RGB+D dataset, these have been collected by the Microsoft Kinect sensor that is described more in subsection 1.1.5. Moreover, the dataset uses three cameras put at the same height but shot from different horizontal angles $(-45°, 0°, 45°)$ for every action. The original paper about the NTU-RGB+D proposes to use two benchmarks for action recognition accuracy. Firstly, a setup called cross-subject (CS) where the person in the training set is different from the person in the validation set. The second setup is cross-view (CV) where instead the camera angle in the training and test set are different from each other. We will throughout the remainder of this thesis refer to the NTU-RGB+D dataset as the NTU dataset.

1.1.5 Ghost body

Throughout the remainder of this thesis a common referred to term is *ghost body*. This is a term coined in this thesis to describe occurrences of skeleton data captured in video samples that are not human. In the NTU dataset the ghost bodies come from oversensitivites in the Microsoft Kinect sensor, which is the tool used to collect skeleton data in the dataset. An example of a ghost body can be seen in Figure 1.1. In action recognition, ghost bodies are most likely not desirable when they are a part of training human action recognition algorithms.

1.2 Problem statement

In a recent implementation of a graph-based action recognition neural net, the authors describe the issue of ghost bodies in the NTU dataset and how they have accounted for them in training of their algorithm [3]. To the best of our knowledge, neither this nor any other graph-based action recognition experiment using the NTU dataset, has explored the impact of the ghost bodies. In addition to not discussing how training might have been affected by ghost bodies, the frequency of them in the NTU dataset has not been presented either. As described in subsection 1.1.4, the NTU dataset is widely used for training and evaluating action recognition algorithms which makes potential issues related to ghost bodies all the more interesting.

1.3 Purpose

The purpose of this thesis will be to explore the impact of ghost bodies in skeleton-based recognition algorithms. The aim will be to answer the following research question:

How frequent are ghost bodies in the NTU dataset, and how is the action recognition accuracy affected by training the model proposed in [3] without the presence of ghost bodies?

1.4 Delimitations

To fully explore the impact of ghost bodies on skeleton-based recognition algorithms, one would need to test several different implementations. However,

due to time and computational limitations, this experiment will only be conducted on one state-of-the-art action recognition model [3]. More specifically an unofficial implementation of the algorithm will be used [11].

As training the algorithm is hardware and time intensive, only one of the two setups in the NTU will be explored - namely the one called cross-subject, CS. This represents a separation of the dataset where the persons differ in training and testing datasets. This is opposed to cross-view, CV , where instead the camera angle in the training and testing data are different.

1.5 Approach

In order to to measure the potential effect on action recognition accuracy caused by ghost bodies in the NTU dataset, an existing state-of-the-art directed graph-based neural network, DGNN [3, 11], will be trained on two different data preparation setups. One where, in simple terms, the ghost bodies are included in training and testing, and one where they are not. The measured difference in action recognition accuracy, if any, between these two setups will serve as the metric of the effect of ghost bodies. A deeper explanation of the experiment and its assumptions are included in section 3.1.

1.6 Outline

In the remainder of this report an overview of the action recognition research field will first be given in chapter 2. A deeper explanation of the experiment and the approach to measure the effect of ghost bodies then follows under chapter 3. Under chapter 4 the outcome of the experiment is presented and explained. Lastly, the significance of the results, potential weaknesses and suggestions on further work is explored in chapter 5.

Chapter 2

Related work

In this section different areas of action recognition will be explored. A brief overview of previous approaches for human action recognition that does not explicitly involve deep learning is first given. Thereafter contemporary deep learning approaches are briefly explained. Lastly, the experiment that serves as the stepping stone for this thesis is explained and explored in depth.

2.1 An overview of human action recognition

Human action recognition is the computer task of interpreting and labelling human action from image or video data [1]. The task is not a trivial one as there are large variations in how human actions take form. Consider for example the task of walking. It may vary in speed, the stride length and even the particularities of a certain person's walking style may differ widely. Many actions also involve more than one person and or interactions with the environment which increase the complexity of recognizing actions even further. Imperfections in the video data, such as occluded body parts is another obstacle. Given these intricacies [1] suggests that a viable human action recognition algorithm should be good at recognizing variations within one class of action as well as be able to accurately make out the differences between different classes of actions.

2.2 Earlier approaches for action classification

The most common method for recognizing actions has historically been and still revolves around extracting features from images in video data in order to assign them an action label [1]. However, there are different approaches to representing the image data as there are many ways to encode and consider the temporal dimension in an action. Given some input data representing an action, the remaining task is to classify it. One type of approach for doing so is called direct classification which means that the temporal information in a video sequence is condensed into a single representation or that the classification is carried out in each frame of a video sequence [1].

2.2.1 Indirect classification

In indirect classification principal component analysis, PCA, has often been used to reduce the high dimensionality of image data [1]. This is done in order to serve as better input data for classifiers in general and action recognition classifiers in particular [1]. Given some condensed form of representation in indirect action classification, one approach for performing the classification involves the algorithm Nearest neighbour classification, NN classification, which in layman's terms compares a distance metric between an input image and a class that the algorithm has been trained on.

The input is then classified as the class it is closest to measured by the distance metric. One issue when using NN classification in action recognition is the high dimensionality of the image data [1]. This has led to implementations where principal component analysis has been used as a precursor to NN classification [12]. The distance metric used in NN action classification has also been subject to research and is important to consider in action recognition using NN [1].

So-called support vector machines, SVMs have also been used in action recognition [1]. In short, SVMs is a way to create vectors that mark the separation of different classes. Relevance vector machine, a probabilistic implementation of an SVM, has specifically been successfully used within action recognition [13] .

2.2.2 Models encoding temporal information

In contrast to direct classification where the temporal information is not explicitly considered there are also so-called Temporal state-space models [1] where different states of a model are connected by probability edges. A state, when this kind of model is used in action recognition, represents the probability of an action at a given time in an action sequence. One implementation of the temporal state-space model is the Hidden Markov Model, HMM. The edges between two states in an HMM represents the probability that an action in one state goes over to another state and as such takes in to account temporal information. In more advanced implementations of HMM for action recognition, a state is not an entire image or body performing an action at a certain time, but rather a certain body part's information at a given time. These approaches have thus used one HMM per body part in a particular action which has proved to simplify the training of the models as there are fewer combinations to consider if only one body part is modelled, compared to all possible states in an HMM that represents an entire body [1].

One weakness of HMMs in action recognition is that it assumes that states are independent over time which is rarely true for human action [1]. As a means to solve this discriminative models are employed, meaning that they are trained to separate different classes rather than learning how to model a certain class [1]. One implementation of such a discriminative model is called conditional random fields, CRF. One study focusing on actions where temporal states of the action are particularly correlated, proved to outperform HMM approaches for classifying actions [14]. However, there is no research to support that CRFs should be more accurate than HMMs for action recognition in general.

2.3 Action recognition using deep learning

In recent years the majority of action recognition algorithms leverages deep learning. A brief overview of the different categories of deep learning approaches within action recognition are outlined in the following section.

2.3.1 Skeleton-based action recognition with RNNs

Previous to using graph neural networks for action recognition, employing recurrent neural networks, RNNs, was the most accurate and widely used approach for skeleton-based action recognition [15]. One reason for how well RNN approaches have worked for action recognition tasks is that it is a vari-

ant of deep learning well suited for learning time-dependencies in sequential data [15]. As action recognition requires interpretation of temporal as well as spatial information, RNNs proves a good tool to use for this purpose. Further extensions of RNN approaches applies so-called long- short-term memory, LSTM, learning.

One approach leverages an LSTM approach to create what they call a Spatio-temporal LSTM algorithm which both models the spatial dependencies of the joints in the skeleton data, as well as the temporal information in the movement between the frames of the video data[15]. Back in 2016 this approach performed on par with contemporary state-of-the-art algorithms and achieved a 69.2 % (cross-subject) recognition accuracy on the NTU dataset which was soon to be outperformed by the advent of CNN action recognition approaches.

2.3.2 Skeleton-based action recognition with CNNs

One alternative approach for using a deep learning in skeleton-based action recognition uses convolutional neural networks, CNNs, to classify and detect action in video data. [3]. A convolutional neural network is otherwise a widely used method for image object detection. When it comes to object detection, it is usually employed on one single image, meaning that CNNs has generally not been used to include temporal information. However, accounting for temporal information is key when it comes to action recognition. To address this issue, CNN approaches used in action recognition represent a skeleton action as a so-called pseudo-image. A pseudo-image in this case refers to compressed skeleton coordinates from several frames in a video sequence. This pseudo-image is then fed to a convolutional neural net for classification. The temporal data of an action is, using this approach thus condensed into a single image [16].

One implementation of this approach was proposed as an alternative to the more common RNN approaches at the time [16]. The authors of [16] achieved an 83.2 % (cross-subject) action recognition accuracy on the NTU dataset which at the time was the state of the art.

2.3.3 Skeleton-based action recognition using graph-based neural nets

The more recent advances in the action recognition field focus on so-called graph-based neural networks, GNNs [3]. The key idea with GNNs lies in how the input data is modelled using a graph. A human skeleton in this approach is represented by vertices and edges where the vertices represent a joint in the human body and where the edges between the joints model human bones [3]. Input data modeled in this way is argued to be a more intuitive approach since the human body can naturally be seen as a graph rather than a sequence or an image as in the RNN and CNN approaches described above. One of the first attempts to apply a directed graph neural netowork, DGNN, for action recognition occurred in 2019 [3]. As that experiment serves as the base for the experiment that will be carried out in this thesis, its contents will now be explored.

Firstly, the experiment argues the superiority of graph-based approaches as it captures both joint and bones data. The authors suggest that for most actions, a movement of a joint (vertex) is highly correlated with the movement of the bone (edge) that is connected to that joint (vertex). Further, measuring and training a neural network to recognize these correlations serves as the base of the authors' experiment. The authors leverage the graph representation in two ways. Firstly, they extract bone and joint information from the video data which is fed into the GNN for feature extraction. Note that this is done on a per-frame basis meaning that the temporal aspect is not accounted for yet. The authors call this the extraction of the spatial information of each frame. Secondly, they use differences between the graph representations in consecutive frames to determine the movement of the graph from one frame to another. This step serves as the temporal dimension or as they call it, the motion information of the action. Lastly, these two input data streams are fed into a two-stream model as its training data.

The key for the model to extracting and capturing the dependencies between joints and vertices lies in its directed graph network-block. The authors make clever use of a directed graph where they firstly use two aggregate functions to pool information from incoming and outgoing edges of a given vertex. Each vertex in the graph is then updated using another function that takes the original vertex and the aggregations from the previous step as its input. Similarly, each edge is updated with a similar function, where the input data instead is the edge and the aggregations from the previous step. The outputs from these two

functions thus have modelled and captured a dependency between the joints and bones which the model can then be trained on [3].

When the experiment was published in 2019, the two-stream approach for the DGNN performed radically better than the contemporary state-of-the-art experiments. On the NTU dataset, an action recognition accuracy of 89.9% was measured (cross-subject).

As this paper serves as the base for the one carried out in our thesis, it is also important to explore how the authors tackled the problem of ghost bodies. Firstly, the authors claim that some of the samples in the NTU dataset contain more than two bodies where some of them are not actually human bodies but rather objects that are wrongly classified in the dataset. These are in fact what is defined as ghost bodies within the confines of this thesis. To deal with the issue of ghost bodies, the authors limit their input data to only contain information about two bodies regardless of how many are actually in the sample. The heuristic for determining which two bodies (if there were more than one in the sample) to extract was to simply look for the two largest sums of joint movement from a presumed body and exclude the rest from the sample [3].

2.3.4 Recent advances in graph-based action recognition

There is more recent research improving on the DGNN [4, 17, 6] where several approaches perform better on the NTU dataset [3]. However, to the best of our knowledge, little or no research has been directed towards analyzing the NTU training data with regards to ghost bodies. Neither has attention been devoted to explore how action recognition could be affected by their presence when training modern action recognition algorithms.

Chapter 3

Method

The Method will present the design of the experiment and its assumptions. A detailed explanation of the approach is also described under section 3.2.

3.1 Experiment design

The approach of the experiment was designed around measuring how the action recognition accuracy is affected by the occurrence of ghost bodies. It is known that the NTU dataset only contains action classes where at most two real human beings are participating. It was therefore assumed that if samples in action classes with only one human in it, contains two or more skeleton data instances, the additional skeletons are likely ghost bodies. It is important to note that the above is an assumption and not a given fact about the data set. However, the NTU is a dataset created in a controlled environment. This suggests that it would be peculiar if the creators of the dataset chose to include other humans than the one performing the action. This strengthens the choice of that simplification. Given this assumption, two different training setups were performed in order to isolate the effect of ghost bodies.

Firstly, the baseline for measuring the effect was to simply train the state-of-the-art DGNN proposed in [3] on the original data, but with only one-person classes. If there were ghost bodies in any of these one-person samples they were thus included when training the algorithm.

Secondly, the action recognition accuracy was compared with that of the second setup. The exact same samples, again only involving one-person classes, was used in training , but with the important difference that the number of

skeletons extracted from each frame in a sample was capped to only include the one with the single most movement across the duration of the video sequence. Since the samples only involved one human, the extracted skeleton in this setup would most likely also be the actual human. This was deemed to be a reasonable assumption but potential weaknesses connected to this will be discussed further in chapter 5.

The reason for limiting the dataset to only look at one-person classes was to simplify the process of isolating the effect of ghost bodies. The input shape of the data in the model depends on the number of expected bodies in the samples. Thus, if the experiment were to include two-person classes, the minimum of extracted skeletons would be two even for one-person classes. Therefore, one-person classes would, in such a baseline setup include both an actual human, and one ghost body. Without manually removing the ghost body skeleton from this one-person action sample, the baseline would contain ghost bodies. Since the purpose of the baseline is to only include actual humans, the issue above would largely complicate the data preparation.

Note that since the data consists of sequences of images, ghost bodies could potentially just have be present in some frames of a sample. A sample in the NTU dataset is a sequence of images and it is therefore possible for a ghost body to be part of some, but not necessarily all frames of it. Thus, in order to measure the frequency of ghost bodies, a sample was defined as a ghost body sample, if a ghost body could be found in at least one frame of it. Lastly, the action recognition accuracies of the two setups described above was compared on the NTU CS (cross-subject) setup.

3.2 Experiment approach

In the following section the procedure for the experiment will be explained.

3.2.1 The model

In order to modify the experiment in [3] existing implementations were researched. As the official implementation was not available, the decision was made to look for alternatives. A well documented unofficial PyTorch implementation was found and used for the entirety of the experiment [11]. This implementation had the entire process documented including data preparation steps, training and evaluation instructions. However, even though the implementation allowed to abstract from the finer details of the model, it is notewor-

thy that this implementation proved to be video memory intense. In effect the model required more then 12 GB of video memory to commence the original training. For our experiments that meant modifying the training process so that it could run on two or more Nvidia TitanX GPUs in parallel.

In order to make sure that this unofficial implementation was in fact a viable replication of the original experiment in [3], the first issue was to ensure that the same action recognition could be measured. As such, the model was trained without any modifications in either the NTU dataset or the model. After 8 days of training on 4 TitanX GPU run in parallel, the training was completed and the testing procedure commenced (which was considerably faster).

3.2.2 Data preparation

The data preparation step was also included in the used implementation of the model. The code for the data preparation was studied and with small modification, every sample used in training and testing could be modified to only train on one body in each sample even though there were several skeletons in it. This could be done due to the fact that each sample frame in the NTU dataset contained a metric for the skeleton count in each frame of the sample, as well as information about whether or not it was a one-person action sample. As such, if first all multi-body actions were filtered out from dataset, but there still proved to be frames containing more than one skeleton, it was considered reasonable to assume that one of those were in fact a ghost body skeleton. To assess this assumption a seven samples were plotted on the original video sequence see Figure 1.1. In the reviewed cases one or more skeletons were in fact found. For example, except for the actual human skeleton, things like chairs and plants constituted the ghost body in question. The frequency and some statistics of the ghost bodies were measured and documented before the training experiment commenced.

3.2.3 Training and testing the model

Given the needed knowledge about the data set, two training sessions were carried out. The first one served as the experiment baseline for comparing the action recognition accuracy. The baseline was constructed so that the original model was trained on all one-person action classes, but without any modifications to filtering out ghost bodies from the samples. In essence, this baseline is exactly how the original experiment was carried out with the notable ex-

ception that only one-person actions are included. This approach was chosen in order to be able to better isolate the effect of training with the presence of ghost bodies. If the original baseline would have been used, it stands to reason that the effects of how well the model classifies two-person actions versus one-person actions would also have been measured. This was outside of the research question and would have weakened the significance the results.

The second training session was also carried out with only one-person classes, but with the important difference that the maximum number of extracted skeletons in each sample frame was limited to one instead of two as in the original experiment. As described under chapter 2, the original experiment extracted (at most) the two skeletons with the most movement over the entire sample as a heuristic to find the relevant human skeletons in the action. Note that this rule was not modified in our experiment - only the number of extracted skeletons were. It is noteworthy that this rule might be too simple, as there is no guarantee that the extracted skeleton is in fact a human skeleton. It could very well, in some cases, be a ghost body. The potential impact of this heuristic is discussed further under chapter 5.

Once both training sessions were finished, testing commenced. Based on limitations of time and computational power, only cross-subject benchmark was performed. The limitations of this is discussed in detail in the Conclusion. Finally, the action recognition accuracy were documented and analyzed.

Chapter 4

Results

Firstly, the frequency of ghost bodies in the NTU dataset are presented. Secondly, the action recognition accuracies from training the DGNN in [3] on the two different data preparation setups are revealed and compared.

4.1 Ghost body occurrences in the NTU dataset

Action class	Count	Share
1 person classes	46231	81.7 %
2 person classes	10347	18.3 %
Total	56578	100 %

Table 4.1: Two-person versus one-person action classes in NTU dataset

Firstly, as seen in Table 4.1, the NTU dataset was found to contain 46231 samples where the action only contains one human. This was opposed to the full dataset with both one-person classes and two-person classes which counted 56578 samples in total. Thus, one-person actions stands for the majority of the data with 81.7% of the samples in the dataset.

Secondly, out of the 46231 one-person class samples used in the NTU dataset 1799 of these were found to contain more than one person in one or more frames of the action sample as seen in Table 4.2. This should, under the assumptions described in chapter 3, be interpreted as the number of samples containing ghost bodies since the action classes, if they do not contain ghost bodies, only contains one single skeleton - namely the actual human. Note

Type	Count	Share
Samples with ghost bodies	1799	3.9 %
Valid samples	44432	95.9 %
Total	46231	100 %

Table 4.2: Ghost body sample statistics for one-person classes in the NTU dataset

again that our definition of a ghost body sample does not require all frames to contain a ghost body. Rather, simply one single frame containing a ghost body will suffice for it to be counted towards ghost body samples even though this is rarely the case. The counting of ghost bodies was explained in detail under chapter 3.

4.2 Action recognition accuracies

Dataset setup (Cross-Subject)	Top 1	Top 5
Original dataset (reported in [3])	89.9%	-
Original dataset (unofficial implementation)	84.13%	96.10%
One-person classes with maximum of two skeletons	84.09%	96.03%
One-person classes with maximum of one skeletons	85.88%	96.93%

Table 4.3: Comparison of action recognition accuracy between different experiment implementations and data preparation setups.

The prediction accuracies presented in Table 4.3 represents the results of the two different data preparation setups. Note again that the accuracies are measured cross-subject (See subsection 1.1.4). The column, *Top 1* represents the action recognition accuracy when the DGNN classifies the correct action as its first choice. The *Top 5* column instead measures the action recognition accuracy when the model recognizes the correct action as one of its top five choices.

As seen in Table 4.3, the action recognition accuracy is slightly higher when ghost bodies are filtered out as there is a difference of 1.79 percentage points in the two different data preparation setups when looking at the top 1 accuracy. The same difference for the top five accuracy measures 0.9 percentage points.

The results of replicating the original experiment in [3] is also noteworthy.

The original paper reports a 89.9 % action recognition, cross-subject on top 1 accuracy whereas the unofficial implementation used in our experiment only achieves 84.13 %. The same measurement on top five accuracy is not reported in [3] and thus cannot be compared. The reasons and implications of the differences between the reported accuracy in the original DGNN and the one measured in the unofficial implementation will be discussed further in chapter 5.

Chapter 5

Conclusion

In the following section the results are discussed and concluded. After that follows suggestions on further research. Finally an answer to the proposed research question is given.

5.1 The significance of ghost bodies in the NTU dataset

As seen in chapter 4, 3.9 % of all one-person action classes were found to contain ghost bodies in the NTU dataset. To answer the research question, it is important to discuss if that is to be considered a significant share.

Firstly, it is worth reiterating that the NTU dataset was created in a controlled environment which means that data quality can be expected to be higher than that of an in-the-wild dataset. In that sense, 3.9 % could be considered to be a significant amount. Furthermore, much of the benefits of using a controlled dataset like this one is to be able to compare different algorithms on different aspects in regards to the action recognition accuracy. The key idea behind comparability in datasets is of course to isolate the performance of the algorithms rather than measure how different experiments prepare the data for its algorithm. However, as seen in [3] the authors have come up with their own way of, at least partially, account for ghost bodies. Other experiments using the same dataset have not mentioned any such data preparation which should suggest that they have not considered it. This could reveal that if different experiments within action recognition use different heuristics for filtering the

data before the actual training, that would partially weaken the comparability of different algorithms.

On the other hand, it can be argued that other benefits of the NTU dataset, such as the cross-view and cross-subject data preparation setups, are such important features for comparing action recognition performance. So much so that noise such as ghost bodies is surmountable. Furthermore, given the rapid development in the field over the past couple of years, recent action recognition research seems to be largely focused on discovering new approaches and algorithms for action recognition rather than fine-tuning data preparation and performance of existing ones. This would point to that, at least for the time being, the occurrence of ghost bodies is not that important for the field and that better data is something to consider when the frequency of breakthroughs within the research field are less frequent.

However, to fully answer if the 3.9 % ghost body frequency is significant, the effect their presence have on the action recognition accuracy must be explored.

5.2 The significance of ghost bodies on action recognition accuracy

As stated in chapter 4, the top 1 (top 5) action recognition accuracy was found to increase by 1.79 (0.9) percentage points when removing ghost bodies from the data. This points to the fact that the presence ghost bodies seem to have an effect on action recognition accuracy. As the top 1 accuracy is (expectedly so) lower than the top 5 accuracies for action recognition in general, the higher increase in top 1 accuracy, when looking at percentage point difference, comes as no surprise.

The effect of ghost bodies could at first be considered quite small but there are number of factors pointing to a 1.79 percentage point increase being a significant difference. Firstly, bare in mind that 3.9 % of the samples contained ghost bodies. In context, this is a rather small frequency that nevertheless seems to have a measurable effect on the action recognition accuracy. It is of course hard to speculate on what the action recognition would be if more samples contained ghost bodies, but most likely it would be significantly lower.

Moreover, efforts to improve action recognition is likely an endeavor of diminishing returns. In simple terms, this suggests that an increase to the accuracy

of 1.79 percentage points when the underlying model already achieves a 80 - 90 % accuracy, is more significant than the same increase where the underlying model performs around a 20-50 % accuracy. The same phenomenon also explains the lower absolute increase of 0.9 percentage points in the top 5 accuracy as this metric is likely subject to even more diminishing improvements than is the top 1 accuracy. It is also worth comparing the 1.79 percentage point increase in action recognition accuracy with that of the improvements seen from introducing new algorithms altogether. As mentioned in chapter 2, contemporary RNN approaches achieved 69.2% in 2016, while CNN algorithms achieved 83.2% in 2017 and lastly DGNNs with 89.9% in 2019. Thus, considering an increase of 6,7 % percentage point increase between modern CNN approaches and a DGNN approach, 1,79 % without changing the algorithm, points to the significance of the results. Note however, that this is not a perfectly comparable metric as these experiments also trained on two-person classes which was not the case for our experiment. The comparison is made merely to point out how much 1.79 % in fact is within this context.

At the very least, it can be concluded that the ghost bodies in the NTU dataset introduces some measure of uncertainty for the DGNN, that makes it worse at recognizing actions. This also points to the 3.9 % ghost body frequency in one-person classes on the NTU dataset to having a significant effect on the accuracy. There are however issues in the experiment that undermine the strengths of the results. These will now be explored.

5.3 Weaknesses

The perhaps largest issue with the results is that the used unofficial implementation was not able to replicate the action recognition accuracy reported in the original experiment [3]. There are a number of factors that could be the explanation for this. According to closed issues reported in the code repository in the unofficial implementation, the author blame hardware limitations, and that certain hyper parameters were missing in the implementation for the measured 84.13 % action recognition accuracy as opposed to the 89.9 % presented in the original paper. This radical difference suggests that no conclusion about the effect of ghost bodies on the DGNN in the original paper [3] can be made. However, it is regardless reasonable to claim that ghost bodies seem to have a measurable effect on the accuracies of directed graph-based action recognition algorithms while not the exact one proposed suggested in [3].

Another issue is that the results only include measurements on the cross-subject

data preparation setup. As mentioned before this was a necessary limitation due to training being resource and time intensive. If the difference in action recognition could be measured on the cross-view data preparation setup, the results would have been considerable stronger.

To truly understand the effect that ghost bodies have, it would also be necessary to look deeper into the dataset and more specifically where the ghost body samples lie. For example, it is possible that there is an uneven distribution of ghost bodies in the training set versus the test set, which could radically affect the measured action recognition accuracy. A deeper analysis of which samples the original model fails to recognize actions on would be needed. It is otherwise possible that none or only a few of these samples actually contain the ghost bodies. If that would be the case, the measured difference in action recognition accuracy would not only or not at all be originating from the occurrences of ghost bodies but rather some other phenomenon.

Lastly, an assumption to simplify measurements was made. As described in section 3.1, any additional bodies in samples with one-person classes were assumed to be ghost bodies rather than another human skeleton. Moreover, when ghost bodies were removed from data samples, the skeleton with the most movement was kept while the others were discarded. The first assumption is reasonable as it is otherwise hard to understand why the creators of the NTU dataset would add humans that are not performing an action to some of their samples. The second assumption is also quite logical if one thinks about what ghost bodies are. If they constitute static objects such as the chair seen in Figure 1.1, the assumption would certainly be considered sensible. In contrast however, it is also possible that there are ghost bodies with moving objects in the background or flickering skeleton joints which would cause the assumption to filter out the wrong skeleton data. A deeper analysis of the samples containing presumed ghost bodies would be needed to give additional support for the assumptions.

5.4 Further research

In order to fully address the proposed research question, the most obvious suggestion on future research regarding ghost bodies would be to to redo the exact same experiment as proposed in this thesis, but where the reported 89.9 % action recognition accuracy can be replicated. While such a replication still cannot be used as a base for a general conclusion about the effect of ghost bodies on graph-based action recognition algorithms, it nevertheless answers

the question that was first laid forth within this thesis.

Except for such a replication as described in the previous paragraph, the search for more general conclusions about ghost bodies in action recognition would be desirable. Some suggestions would be to replicate this experiment for several state-of-the-art graph-based algorithms. As testing the cross-view data preparation setup was omitted from this experiment, measuring the results on this setup would also be a suggestion on how to further generalize the results.

Lastly, despite all weaknesses clearly diminishing the results, the fact remains that some phenomenon correlated with the ghost bodies seems to have been found here. This points to the last suggestion on further research, which would be to consider working on improvements on the NTU dataset. A version of the dataset, rid of ghost bodies, could be created as a benchmark of how well algorithms deal with the presence ghost bodies in training data. Given the importance of this dataset within the field of action recognition, such an endeavour might be worthwhile in order to achieve better comparisons for present and future action recognition algorithms.

5.5 Summary

In conclusion, 3.9 % of one-person class samples contain ghost bodies. Moreover, there is a measurable increase in action recognition accuracy when training a DGNN without the presence of ghost bodies. However, there are several weaknesses in the experiment that needs to be addressed before a definitive answer to the research question could be given. As the proposed question was to explore the effects on the DGNN proposed in [3], the most important drawback to the results is that the experiment could not be fully replicated. Despite this, findings suggests that the presence of ghost bodies in the NTU dataset decreases action recognition accuracy for a graph-based action recognition algorithm by 1.79 %. Given how widely used the NTU dataset is within the field, further exploration of ghost bodies and their effects is both interesting and important.

Bibliography

[1] Ronald Poppe. "A survey on vision-based human action recognition". In: *Image and Vision Computing* (2010).

[2] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. "Going deeper into action recognition: A survey". In: *Image and Vision Computing* (2017).

[3] Lei Shi, Yifan Zhang amd Jian Cheng, and Hanqing Lu. "Skeleton-based action recognition with directed graph neural networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

[4] Ziyu Liu et al. "Disentangling and unifying graph convolutions for skeleton-based action recognition". In: *CoRR* (2020).

[5] D. Q. Huynh L. Wang and P. Koniusz. "A Comparative Review of Recent Kinect-Based Action Recognition Algorithms". In: *IEEE Transactions on Image Processing* (2020).

[6] Ke Cheng et al. "Skeleton-based action recognition with shift graph convolutional network". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 183–192.

[7] Amir Shahroudy et al. "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis". In: *CoRR* (2016).

[8] Bo Li et al. "Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN". In: *IEEE International Conference on Multimedia Expo Workshops (ICMEW)* (2017).

[9] Tom M. Mitchell. *Machine Learning*. 1997.

[10] W.J. (Chris) Zhang et al. "On Definition of Deep Learning". In: *2018 World Automation Congress* (2018).

[11] Ziyu Liu (kenziyuliu). *Unofficial DGNN-PyTorch*. github.com, 2018. (Visited on 2021).

[12] Wang et al. “Learning and Matching of Dynamic Shape Manifolds for Human Action Recognition”. In: *IEEE Transactions on Image Processing* (2007).

[13] Antonios Oikonomopoulos, Ioannis Patras, and Maja Pantic. “Spatiotemporal salient points for visual recognition of human actions”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* (2006).

[14] M. Ángeles Mendoza and Nicolás Pérez de la Blanca. “Applying Space State Models in Human Action Recognition: A Comparative Study”. In: *International Conference on Articulated Motion and Deformable Objects* (2008).

[15] Jun Liu et al. “Spatio-Temporal LSTM with Trust Gates for 3d Human Action Recognition.” In: *CoRR* (2016).

[16] Chao Li et al. “Skeleton-based Action Recognition with Convolutional Neural Networks”. In: *CoRR* (2017).

[17] Maosen Li et al. “Dynamic Multiscale Graph Neural Networks for 3D Skeleton Based Human Motion Prediction”. In: *CoRR* (2020).

TRITA-EECS-EX-2021:483