Master thesis
Computing Science

Radboud University

Supporting Sign Language Learning With a Visual Dictionary

Author:
Mark Wijkhuizen
S4659147

First supervisor:
Prof. M.A. Larson
m.larson@cs.ru.nl

Co-supervisor:
prof. dr. O.A. Crasborn
o.crasborn@let.ru.nl

September 30, 2021

Contents

Abstract

Objective
This research explores the linguistic and technological challenges of developing a visually searchable sign language dictionary that allows sign language learners to search by means of the Sign Language of the Netherlands (NGT), instead of using Dutch translations of signs or sign properties. All code is made publicly available on GitHub [1].

Methodology
Interviews were held with stakeholders in the deaf community to determine requirements for a visually searchable dictionary Additionally, the interviews served to gain an understanding of how sign language learners perceive similarity between signs in order to quantify sign similarity. This quantification served to assess the quality of search results. Consequently, a design and creation approach was used to test the suitability of Motion Fused Frames (MFFs) in an Information Retrieval (IR) setting and to introduce a key frame extraction method for sign recordings based on sign structure.

Results
Key requirements identified during the interviews are: the ability to retrieve semantically similar signs, the ability to view signs from the side, iconicity explanation, and information on the context signs can be used in. The established metric for sign similarity is defined as: the mean of the in this research introduced similarity quantification for location, movement, handedness, and hand shape. As a result, the dictionary's performance can be evaluated by applying the sign similarity as relevance in the $NDCG$ metric. MFFs perform consistently better than RGB frames with a $top20Accuracy$ improvement from 36% to 44% and an $NDCG@20$ improvement from 55% to 58%. The developed key frame extraction method showed an $top1Accuracy$ improvement from 44% to 55% and an $NDCG@20$ improvement from 58% to 59%.

Conclusion
The developed model which relies on MFFs and introduces a key frame extraction method for sign recordings provides a suitable data representation in an IR setting with highly discriminative value. It is a first step in the introduction of a visually searchable sign language dictionary. This research contributed to the field of IR and linguistics by expanding the knowledge on user requirements, sign similarity, MFFs data representation and domain specific key frame extraction.

Executive Summary

This research explores the linguistic and technological challenges of a visually searchable sign language dictionary that allows sign language learners to search in Sign Language of the Netherlands (NGT), instead of Dutch translations of signs or sign properties. There is no official written form for sign language, preventing the development of sign language equivalents of spoken language dictionaries where user can search in textual notations of NGT. Moreover, the unofficial written forms are primarily used in academic environments and are generally considered too complex for public use [2]. The absence of a written form could be found in the parallel structure of phonemes in sign language, compared to the sequential structure in spoken language. Spoken languages have a single dimensionality of phonemes, sound, whereas sign language have multiple dimensionalities of phonemes, such as location, movement and hand shape [3]. Since text also operates on a single dimension it is challenging to capture the multiple phoneme dimensions of signs in text. A visual dictionary would not require a written form and would also allow sign language learners to search in NGT rather than in Dutch or on sign properties. This information retrieval (IR) approach allows sign language learners to retrieve the most similar signs for a given input sign. Similar signs are also being retrieved when the input sign is articulated improperly, as can be expected with beginning signers.

Current research on sign language recognition and IR has limitations on a variety of aspects. Firstly, visual sign language dictionaries have thus far only been constructed in a experimental setting [4][5]. The requirements sign language learners would have from a publicly available visual dictionary are thus unknown. Secondly, sign similarity has only been manually quantified, which is not scalable and subjective [5]. Evaluating the relevance of results in an IR setting is therefore time consuming and inaccurate. Quantification of sign similarity from a sign language learners point of view would automate the evaluation process. Thirdly, deep learning methods have thus far only been applied to sign language classification, whereas non-deep learning methods are applied to sign language IR [6][7][8][9]. A deep learning sign IR system could extend the knowledge on sign classification to the underexplored field of sign IR. Lastly, many generic key frame extraction tools developed over the years, but they do not solely take into account hand movement, but any movement [10] [11]. OpenPose, a body keypoint detection tool, would address this shortcoming by isolating hand movement, resulting in a domain specific sign key frame extraction tool [12]. This results in more discriminative frames and could improve the performance of the visual dictionary.

Identifying the requirements and quantifying sign similarity was addressed by performing interviews with sign language students, interpreters, teachers and PhD candidates. Requirements identified during interviews were, first of all, the preference to use the tool as a dictionary, rather than a tool to search for similar signs. Next, the functionality to view semantically similar signs would support students with expanding their vocabulary in the thematic field they are currently exercising. Thirdly, the option to view signs from the side would clarify movement in the z-axis. This functionality is desired, because videos of signs are typically recorded from the front, making movement in the z-axis is hard to interpret. Fourthly, Iconicity explanation would give insights in the origin of a sign, making the sign more intuitive and thereby easier to remember. Furthermore, giving examples of correct and incorrect context usages of a sign would clarify to sign language learners a better understanding of the semantics of a sign, which is crucial since some Dutch homonyms have different signs for each meaning. Lastly, design requirements consists of a minimal user interface and clear instructions on how to record the input video. These requirements were formulated as user stories and should provide clear insights in the wishes of sign language learners for an innovation that has not come to market yet. Requirements should ensure developers to create a tool which meets the demands of sign language learners.

After clarifying the requirements for the visual dictionary an evaluation metric needed to be constructed to evaluate the relevance of retrieved signs from a sign language learner's perspective. This does however require a relevance score for each retrieved sign to the input sign. This led to

introducing a quantification method for sign similarity, which takes the input sign and the retrieved sign and gives a similarity score in the range $[0, 1]$. During interviews sign similarity was extensively discussed and resulted in a quantification method based on the four phonemes location, movement, handedness and hand shape. Locations are divided into regions in which different locations are easily confused. These locations are upper and lower head, upper, lower and sub torso, neutral space, weak hand and arm. Movement is characterised by the axis it operates on, independent from the direction. Movement in the z-axis is considered less visually dominant than horizontal and vertical movement. Handedness is divided in one and two handed signs, where two handed signs are further subdivided. Location, movement and handedness are structured as a taxonomy tree where similarity between is defined as the node distance divided by the maximum distance in the tree. Generic hand shape properties are based on the work of Van der Kooij, because interviewees had difficulty defining generic hand shape properties [3]. Similarity between hand shapes is defined as the fraction of overlapping properties where similarity score for both the strong and weak hand is computed. These five similarity scores are summed and divided by five to result in a final similarity score based on location, movement, handedness and hand shape. This similarity score is used to evaluate the relevance of the retrieved signs for sign language learners using the widely used normalised discounted cumulative gain [5] [13].

With the evaluation metric defined the actual visual dictionary could be developed. The work of Kopuklu et al. is used as a basis [14]. This model introduced Motion Fused Frames (MFFs), where optical flow frames are appended to RGB (Red, Green, Blue) images to both capture spatial and temporal features. This data representation is selected since signs are characterised by both spatial properties, such as location and hand shape, and temporal movement properties. To evaluate the performance a validation and test is created of respectively 100 and 153 different signs recorded in homely setting with conventional webcams to closely imitate usage in practice. Compared to RGB frames, MFFs improved the $top20Acc$ from 36 to 44 and the $NDCG@20$ from 55 to 58 on the test set. This indicates the appended optical flow frames containing temporal information yield added value and improve the relevance of results for sign language learners. When varying the number of optical flow frames no consistent improvement is perceived, making the added value of additional optical flow frames unknown.

Lastly, sign key frame extraction method for sign language is proposed. Signs are characterised by the starting location and hand shape, optional movement and end location and hand shape [15]. With OpenPose the left and right wrists are detected in each frame. By computing the location coordinate deltas between each pair of consecutive locations, movement can be quantified for both the x and y axis. The dominant hand is selected by choosing the hand with the maximum total movement. The first key frame selected is the first static frame after the initial vertical movement from the neutral position, thereby capturing the starting location and hand shape. The end frame is the first static frame before the last vertical movement to the neutral position. This frame should capture the final position and hand shape. These near static start and end frame should have minimal motion blur and should therefore capture the spatial phonemes location and hand shape. Optional movements are identified by finding local optima between the start and end frame. The frame after the local optima is selected to capture the maximum movement in the optical flow frames. This domain specific sign key frame extraction method should capture both spatial and temporal information and thereby providing information on the location, movement and hand shape. The method resulted in a $top1Acc@20$ improvement from 44 to 55 and an $NDCG@20$ improvement from 58 to 59. This performance improvement indicates the added value over uniform sampling. This key frame extraction method could function as a baseline for further research on domain specific key frame extraction.

This research explored the linguistic and technological challenges of a visual NGT dictionary. The identified requirements and constructed quantification method for perceived sign similarity by beginning signers should optimise the visual dictionary for the needs of sign language learners. Empirically substantiate the added value of temporal information in MFFs and the introducing a key frame extraction method for sign language should further enhance the performance of a visual sign dictionary. These findings should provide fundamental knowledge to get a step closer to the introduction of a visual searchable NGT dictionary that supports sign language learning.

Acknowledgements

I want to thank Martha Larson for helping me by providing with thesis options, from which this thesis was chosen, the daily supervision and extensive discussions. I want to thank Onno Crasborn for introducing me to the sign language field and hosting the thesis. For the actual finalisation of this thesis I want to thank Jelle, Dylan and Sam for providing extensive feedback in the last couple of weeks.

Chapter 1

Introduction

There is increasing attention for Sign Language of the Netherlands (NGT) in Dutch society. In the autumn of 2020, the Dutch parliament unanimously accepted a law that formally recognises NGT as a language of the Netherlands [16]. This was a major milestone for the approximately 60.000 people who use NGT [17]. An increasing number of people is interested in learning NGT as a second language, which could help create a more inclusive society for deaf and hard-of-hearing people.

Despite increasing interest in NGT, available supportive technologies to learn NGT are rather limited. This lack of supportive tools results in missed learning potential, as people trying to learn NGT cannot receive the optimal support. Currently existing tools allow sign language learners to search signs using a Dutch to NGT dictionary or an online tool to search signs using sign parameters [18] [19]. None of the current tool allows beginning signers to search signs using their native language, NGT. This would require a visually searchable sign information retrieval system, of which the linguistic and technological challenges will be explored in this thesis.

1.1 Fundamental complexities of sign languages

The lack of supportive technologies for people that are learning NGT can be partially explained by the difficulties that are specific to signed languages, as opposed to spoken language. Spoken languages have official written notation. The International Phonetics Association has even made a universal phonetical alphabet to notate any spoken language [20]. Signed language on the other hand lack a comparable official written notation. This makes it challenging to create dictionaries for signed languages. Signed languages have unofficial written notations [2]. However, these unofficial written notations are generally considered to be too complex for sign representation in dictionaries and are primarily designed and used in academic environments [2]. This complexity can mainly be attributed to a difference in the amount and ordering of phonemes; the smallest element in a language that can change the meaning of a word. Signs have multiple dimensionalities of phonemes, i.e. movement, hand shape and location, all operating simultaneously. In contrast, spoken languages have a single phoneme dimension, consisting of consonants and vowels that are linearly sequenced [3]. Written text operates at the same single dimension as spoken languages, since words can be viewed as a mapping of sound to a string of characters, where the character position indicates the sequential ordering. Signs can include a nonmanual aspect, i.e. facial expressions, which could be considered a fourth dimensionality of phonemes in signed languages.

The difficulties of notating signs using symbols make it tempting to use diagrams, as this captures both the spatial and facial features. However, such notation lacks a sequential structure. A sequential structure could be added with a series of diagrams, but this would be spatially inefficient. Another option is adding arrows indicating the movement. This, however, would make the diagrams harder to interpret. The complexities inherent to signed languages make the development of an official written notation challenging. This makes it difficult to create dictionary with a signed language as source language, because there is no textual notation to use as index. In turn, this leads supportive tools for sign language learners, such as dictionaries, to lag behind in functionality.

1.2 Shortcomings of current sign language technologies

Current NGT dictionaries consist of a Dutch to NGT dictionary called Basiswoordenboek Nederlandse Gebarentaal (Base dictionary NGT) [19]. The sign notation is a single diagram, which eliminates the sequential structure of signs and making the dictionary spatially inefficient. The dictionary consists of roughly 3,000 words, in comparison to 50,000 words in the Van Dale Basiswoordenboek Nederlands ('Basic Dictionary of the Dutch') [2]. This dictionary allows users to look up the NGT translation of a Dutch sign. There is however no dictionary from NGT to Dutch, which would allow beginning signers to look up the meaning of signs.

A search tool publicly accessible on the website of Nederlands Gebarencentrum ('Dutch Sign Centre') allows users to search NGT signs using parameters, such as the location and hand shape [18]. This is, arguably, the closest tool to a NGT to Dutch dictionary available. The tool does however require the user to have extensive knowledge about sign notation and eliminates the sequential structures of signs by purely focusing on the spatial features [2].

Moreover, current tools do not allow for the retrieval of signs that are visually similar to an input gesture, only exact matches are retrieved. Most phonological errors happen by small mistakes in either the configuration of the arm, position and especially the action itself [21]. Beginning signers could find it challenging finding signs, since it would require them knowing the exact sign properties.

Currently, no tool exists which allows users to search in NGT, only to NGT, which forms a key problem for learning NGT. For example, when learning English, the lack of both an English dictionary and a dictionary that translates English to the native language would cause a major obstacle. This research will address this lack of a dictionary searchable in NGT by exploring the linguistic and technological challenges involved with creating a visual NGT dictionary that support sign language learning.

To tackle the technological challenges of a visually searchable dictionary, concepts from both computer vision and information retrieval are combined. This technology is based on insight in sign language learning and sign similarity, obtained from interviews with stakeholders. Visual sign language recognition is a subdomain in computer vision that classifies images and videos of signs. In recent years, there has been an increase in research on sign language recognition, driven by the introduction of deep learning [22]. Applying current research on a visual dictionary to support sign language learning results in four unexplored challenges addressed in this research.

1.2.1 Practical applications

First, the practical application of sign language recognition systems is often missing. Many impressive results have been achieved using complex deep learning systems [6][7][8], but without practical applications the societal value of these results is limited. Especially the design of such a system is often left out, because the current goal of research is to optimise an evaluation metric. The focus is thus currently on improving the performance, rather than on creating a system that brings societal value. A system should contain all desired functionalities and have an intuitive design, otherwise the societal adoption will be minimal. Input from users and stakeholders is critical to identify the desired functionality and design. Requirements from sign language learners should therefore be elicited to ensure a match between the implemented and desired functionality.

1.2.2 Evaluating sign IR systems

Second, an evaluation metric that quantifies the relevance of the retrieved signs to a sign language learners should be determined. An evaluation metric quantifies the performance, which can be used to compare different configuration of the same IR system and to compare different IR systems. The dataset currently only provides class labels, which could only be used in classification metrics. To evaluate IR systems, typically relevance labels are required. These labels quantify the relevance of a retrieved document to an input query [13]. The relevance to sign language learners should therefore not rely on class labels, but on perceived similarity between the input sign and retrieved signs instead. This requires a quantifying method for sign similarity. This is not a trivial task and

will require extensive insights in the mistakes sign language learners make and structure of signs. An appropriate evaluation metric will, for a given input sign, reward retrieved signs with a high similarity to the input sign, whilst also take into account the order of the results.

1.2.3 Sign information retrieval

Third, current research focuses on sign language classification, leaving out information retrieval systems [15]. In comparison to sign language classification, information retrieval systems map a query to a set of similar signs, ranked on similarity. Instead of having a single result, an IR system thus has multiple ordered results. In an IR setting the amount and order of the results bring extra dimensions to the system, which are not addressed with sign classification. Many deep learning sign classification systems have been made, among which a model using MFF's (Motion Fused Frames) [14]. Here RGB (Red Green Blue) and optical flow frames are fused to a single image containing both spatial and temporal information are trained using a CNN (Convolutional Neural Network) to classify videos of people performing signs in front of a webcam. The MFF data format allows a frame to contain both spatial sign information, such as location and hand shape, and the temporal movement information.

Empirically evaluating the applicability of MFFs to a IR setting will provide insights in the dominance of MFFs over RGB frames in other domains. Furthermore, the sign similarity evaluation metric will show whether this data format captures the features desired by sign language learners. Exploring the performance of MFFs in IR setting would further substantiate the applicability of MFFs to sign language.

1.2.4 Key Frame Extraction

Finally, the potential of a domain specific key frame extraction method for sign language videos is unexplored. Both images and videos are being used as input in sign language recognition [23][14]. In contrast to image input, video input conventionally requires a fixed number of frames to be selected, this frame selection process is called key frame extraction. Existing key frame extraction methods typically do not discriminate between sources of motion and capture any movement, independent from semantics [24][25]. For example body movement and movement in the background thus be marked as salient frames.

Noise in low quality webcam videos could introduce another challenge. Noise could also be interpreted as salient, since noise changes the pixel values. Conventional key frame extraction methods are developed and benchmarked on high quality videos, making their performance on low quality webcam videos unknown [10] [26] [27].

Not discriminating between sources of motion in noisy videos could reduce the discriminative value of the extracted frames which. This, in turn, reduces the quality of the document representation which would reduce the performance of the IR model. Isolating hand movement would prevent detection of irrelevant movement by filtering the movement on semantic origin, which would not happen with conventional key frame selection.

A domain specific sign language key frame extraction method based on hand movement would therefore allow for semantically aware key frame extraction. This would increase the discriminative value of selected frames, thereby increasing the discriminative value of the document representation. Subsequently, this would improve the performance of software systems that process recordings of signs, such as visual sign dictionaries.

This research will address all four problems by introducing a visually searchable sign dictionary for sign language learners.

1.3 Research Goal

The goal of this research is to create a visually searchable NGT dictionary in order to support sign language learning. This requires getting insights in the desired functionality, and evaluation of a system that supports sign language learning. Next, technical challenges such a system presents need

to be explored. All code developed during this thesis is made publicly available on GitHub [1]. To achieve this goal an IR system will be designed, which allows sign language learners to retrieve identical and similar signs with visual search queries.

1.4 Research Questions

The research objective will be addressed with the following overarching research question and sub questions.

- How can the learning potential of a visual sign language dictionary be maximized?
    1 What requirements do users have for an IR system that supports sign language learning?
    2 How can the relevance of results for sign language learners from a sign language IR system be evaluated?
    3 How do MFF compare to RGB frames as document representation in an IR system?
    4 How does key frame extraction using body keypoints compare to uniform sampling?

These research questions have the following dependencies, as visualised in figure 1.1. RQ 1 will answer which requirements sign language learners have from the visual dictionary and how to sign similarity is perceived. RQ 2 will use the obtained insights on sign similarity to quantify sign similarity and construct an evaluation metric. RQ 3 and 4 are dependent on RQ 2, because the comparison will be based on the evaluation metric constructed RQ 2.

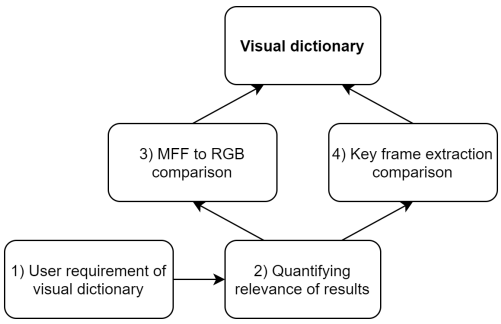

Figure 1.1: Dependencies between the four sub research questions

This research will contribute to both the machine learning and sign language research field. The main contributions in the machine learning field consists of insights in challenges concerning the adaptation of a sign classification system to a sign IR system. Moreover, the potential of MFF's (Motion Fused Frames) and different key frame extraction methods in an IR setting are explored. The contribution to the sign language research field consists of requirements for a visual dictionary and insights in perceived sign similarity by sign language learners. By combining these two fields the theoretical knowledge on sign recognition is applied to a societal problem, the lack of supportive tools to learn sign language. This explores the challenges encountered when introducing theoretical knowledge to real world applications.

1.5 Host Organization

This thesis is performed at the Centre for Language Studies (CLS) research institute, which is part of the Faculty of Arts at the Radboud University in Nijmegen. It belongs to the Centre for Language and Speech Technology (CLST) research department. Where the CLS has a broad focus on the understanding of the nature of language and the way it functions, the CLST has a more narrow focus on the technological application supporting language and speech technology.

CLST's vision is to ensure that no citizen is excluded from the constantly evolving information society we live in, especially the disadvantaged, such as the deaf. They do so to make sure that everyone – the elderly, the young, natives, non-natives, those highly educated, those lowly educated, people with communication disorders, etc. – has access to this data and can (continue to) participate in our information society. CLST joins forces to focus on the following research areas of knowledge extraction,communication in health care, resources & tools and language Learning and teaching [28]. This research will combine the latter two research areas through the creation of a system that

supports sign language learning by allowing learners to visually search for signs and phonetically similar signs.

1.6 Thesis outline

This thesis is structured as follows. Firstly the theoretical background is discussed in chapter 2 followed by listing the conceptual frameworks used in this research in chapter 3. Next, the methodology is explained in chapter 4 and the setup is elaborated in chapter 5. The results of the four research question are described in chapter 6, chapter 7, chapter 8 and chapter 9. The thesis is finalised with a discussion and conclusion.

Chapter 2

Theoretical background

This chapter explorer available research on topics surrounding the four research questions to get a sound theoretical background. Firstly, literature on requirements engineering and user stories is consulted for RQ1. Next, RQ2 requires literature on sign language structure and learning, as well as comparisons between spoken and signed languages. Thirdly, for RQ3 both IR (information retrieval) and MFFs (Motion Fused Frames) are discussed. Lastly, literature on key frame extraction and pose estimation is consulted for RQ4.

2.1 Requirements Engineering

RE (Requirements engineering) is a software development paradigm which can be defined as "the branch of software engineering concerned with the real world goals for functions of and constrains on the software systems" [29]. Less formally, RE can be seen as a software engineering method concerned with defining the goals of a software system and the imposed restrictions. A requirement can be broadly defined as"anything that drives design choices" [30]. These can be functional requirements such as the ability to view certain information, but can also involve non-functional requirements, such as security or hardware constraints [31].

Due to time constraints, not all types of requirements can be addressed. The focus of this thesis will therefore be primarily on the functional requirements concerning what stakeholders expect the software to do. To a lesser extent, the external interfaces and performance will be covered. Interfaces and performance concerns the interaction between users and hardware, hardware internally and speed and response time of the system. Attributes and design constraints imposed on an implementation will be left out of scope. Attributes concern portability, correctness and maintainability, whereas design constraints imposed on an implementation involve any required standards, implementation language, resource limits and operating environment [32].

RE starts with eliciting requirements from stakeholders before software development begins. Visual progress is frequently shown to stakeholders during the development process. Requirements can be added, modified or removed after each iteration [33]. This research will only focus on the first step, getting the initial requirements for a visually searchable dictionary for sign language learners. These requirements will be formulated in a widely used user story format [34]. These user stories formulate a requirement as a functional wish from a users perspective, the most well-know format popularised by Mike Cohn is: *"As a ⟨type of user⟩ , I want ⟨goal⟩, so that ⟨some reason⟩"* [35]. User stories' quality can be assessed with a framework covering syntactic, semantic and pragmatic properties [34]. This requires user stories to for example be atomic, unambiguous and unique.

A paper by Malviya et al. analyses the RE eliciting process [31]. Key findings were the anchoring of stakeholders to examples, indicating questions should be formulated broadly to stimulate creative thinking. Stakeholders could otherwise stick to provided examples or implied solutions in questions, preventing their own idea generating process.

Moreover, the extensive IEEE (Institute of Electrical and Electronics Engineers) recommended practice for SRS (software requirements specifications) document will function as a baseline when

reporting the requirements by using their proposed template [32]. This template lists the required sections of a complete RE report.

An important goal of RE is to reduce the development effort by forcing all stakeholders to consider all requirements before design and implementation begins [32]. This extensive focus on correct requirements eventually pays off, because correcting wrongly implemented requirements is more costly than correcting a requirement [36]. Correction costs even increase exponentially throughout the development cycle, which is on a spectrum of correcting a textual requirement to modifying an operational software system. In addition, 40-60 percent of all defects found in a software project are accounted by errors made during the requirements phase, making it a critical phase to successfully finish a software project [37]. All these problems and costs can be linked to a expectation gap: "a difference between what developers think they are supposed to build and what customers really need" [32].

Especially with novelties, such as an visually searchable sign IR system, the customer needs are unknown and unexplored, because such a system has not been commercially available yet. Such project is called a green-field project and has no knowledge based on prior work, in contrast to a (COTS) commercial off-the-shelf software solution [33].

2.2 Sign Language

For both sign language learning and spoken language learning the origin of the mistakes can be typically found in the phonology rather than the semantics [21]. With spoken languages phonetically close words like "cat" and "bat" could be confused, rather than the semantically close words "cat" and "paws". In signed languages this same phenomenon is present, however the phonetics are visual, rather than vocal. This indicates signs are remembered by signers in terms of simultaneous independent formational parameters, which can therefore be independently pronounced correctly or incorrectly [21].

Similarity is therefore present in the pronunciation. The phonetic errors made by sign language learners can be typically found in the hand configuration and movement, sometimes in the forearm configuration, but no errors were made solely in the location during the experiment [21]. These findings indicate sign language learners pick up the location fast, whilst having difficulties with the hand configuration. When looking up signs in a visually searchable dictionary users are therefore more likely to correctly articulate the position and forearm position of unfamiliar signs, but not the hand configuration.

Another difference between mistakes made in spoken and signed languages can be found in the parallel nature of signed languages, in contrast to be sequential nature of spoken languages [38]. Whereas spoken words consist of a list of syllables, signed words consist of an action: a list of states. Those states can be defined by Stokoe's formational division and therefore have multiple actions in parallel. A sign can be performed at the correct position, but with a wrong hand shape. This parallel nature of signs gives an extra dimension to sign language learning, because students need to focus on multiple aspects simultaneously, in contrast to spoken language. The formational division and parallel nature of sign language forms a theoretical starting point when designing the interviews on the didactic value of the retrieved signs.

2.3 Information Retrieval

As described in section section 2.2, signs are actions and therefore exist of a list of states. These states can be digitally captured with a video recorder to be processed by a computer. This capturing can be done with an RGB (Red, Green, Blue), depth or thermal camera to capture different properties of the action [22]. RGB videos can be further processed to create skeletal videos to solely capture body movement and posture [22]. These input modalities can be combined to improve the performance of these models, which makes them more complex and harder to train [7][6]. These modalities also focus purely on the spatial properties of an action, not the temporal. Motion information can however be added with optical flow frames. Kopuklu et al. proposed a data fusion method where

optical flow frames are appended to RGB frames, thereby fusing temporal and spatial information in a single image called a MFF (Motion Fused Frame) [14]. MFF have thus far only been used in the context sign language classification [14]. It is therefore unknown how this technique performs in an IR setting.

Sign language IR systems have been developed before using non-deep learning approaches. DTW (dynamic time warping) has successfully been applied to a dataset of 1113 signs with a $top10Acc$ of 78% [39]. This system did however require users to manually mark the start and end of the sign in the video, indicate the handedness of the sign. Moreover, a hand detection system needs to process each frame followed by manual verification and improvement. Since the videos were recorded in a studio setting the generalisability of the results to real world setting is questionable. A deep learning approach without human assistance would streamline the IR process. A recently published paper by M. Fragkiadakis and P. van der Putten achieved an $accuracy$@20 of 71% using DTW with a 1200 sign lexicon. Despite this method not relying on user interference the input video still needs to be processed for body key point detection, a deep learning approach would eliminate this need.

There is a wide variety of metrics for IR applications that measure the relevance of the retrieved results, such as $NDCG@K$, $MAP@K$, $Precision@K$ [13]. These metrics aim to quantify the relevance of the first K retrieved documents to the user given a discrete relevance score for each retrieved document to the query. The metrics require documents to be labelled, which is a subjective and time consuming process. Where MAP and $Precision@K$ require samples to be labelled either relevant or not-relevant $NDCG$ requires samples to have a relevance score. Similarity between signs will be a spectrum from dissimilar to similar, and therefore a continuous relevance score instead of a binary relevance label. Another option is to rate the results with an expert group, where a group of information specialists rate the results for given a query [40]. This is time consuming and need to be repeated for each iteration of the IR system. This validation method does however best reflect real world performance of an IR system. For development iteration this method is not suitable, as many iterations and even hyper parameter searches will be required. Therefore, the distance between signs needs to be quantified to appropriately evaluate the IR system in a feasible time frame.

Lastly, humans can detect and learn patterns from a few examples, deep learning models typically require many samples before generalising adequately [41]. When only a few samples per class are available, this is called few-shot learning. The provided dataset contains only one sample per class, a special few-shot learning case called one-shot learning with only one sample per class available for training.

Few-shot learning performance is affected, firstly, by the hypothesis space H, constrained by the used model, and secondly, by the number of samples I in D_{train} [41]. Approaches affecting H consist of parameter sharing, where two similar tasks share the first few layers with private output layers. This allows the smaller dataset to learn generic information with assistance of the larger dataset in the shared layers. This method requires only the private output layers to be trained solely on the smaller dataset. Another popular method is transfer learning, where a model is pre-trained on a similar task with a large dataset to converge the random initialised parameters $h \in H$ closer to the optimal $\hat{h}$ [41]. Data approaches aim to increase the number of samples I to $\tilde{I}|\tilde{I} > I$. Conventional techniques consist of horizontal/vertical flips/shifts, cropping, padding and rotating. In recent year more advanced methods have been introduced, such as masking methods like GridMask where parts of an image are obscured and mixing methods like CutMix where parts of two images are combined into one image [42][43]. This wide variety of few-shot learning approaches gives confidence in achieving competitive performance using deep learning.

2.4 Key Frame Extraction

signs are actions, which are represented as a list of states when recorded with a camera. These videos can differ in captured frames per second (FPS) and length, resulting in a wide variety of total frames. In addition, a recording of a sign can contain padding at the start and end with no user movement. Typically, a machine learning model will require an input with a fixed amount of frames [7][6][14]. Selecting the most salient frames will result in the maximum performance, however this

is a complex task. A baseline key frame extraction method is uniform sampling, selecting frames on a fixed interval [44]. This method does not use the content of the video, but is computationally efficient.

More advanced methods rely on stylistic information and select frames based on temporal and spatial saliency [26]. Here temporal saliency indicates the difference between frames and spatial saliency indicates how feature rich an individual frame is. Other stylistic methods rely on clustering HSV (Hue, Saturation, Value) histograms using a k-means algorithm [10]. Optical flow, which estimates the displacement field between two frames, has also been applied to determine key frames [45][11]. Here the displacement field for each frame has been summed to construct a graph indicating the temporal saliency where local minimums were assigned as key frames.

In signs, the salient frames are those which capture the sign properties which together make a sign unique, which will be further discussed in section 3.3. Subtle movement of the body or any movement in the background is not of importance, but will typically be marked as salient with conventional key frame extraction methods. In recent years, with the introduction of deep learning, a tool called OpenPose has been developed to detect body, hand and face keypoints in images [10]. OpenPose could be applied to detect the coordinates of hands in a frame and quantify hand movement by comparing coordinates between frames. A key frame extraction method based on OpenPose could therefore solely rely on hand movement. This would make the key frame extraction method semantically aware of the object that causes saliency detection. This method would result in a domain specific key frame extraction tool for videos of signs.

Chapter 3

Conceptual framework

This chapter presents the conceptual frameworks that support this research. In section 3.1 the IR framework used to design the visual dictionary is discussed. The next section discusses motion fused frames, which function as a data representation of sign videos. Lastly, a formational division of sign language morphemes is presented. This formational division supports the design of the interviews covering sign similarity and describes the sign properties used for constructing the domain specific key frame extraction method.

3.1 Information Retrieval

Information retrieval systems map a user query to a ranked list of matching documents in a collection. The information retrieval process by Lancaster and Warner shown in figure 3.1 is used as a framework when creating the sign language information retrieval system [40]. This framework is chosen because it gives a generic outline of the full IR process from information need/problem to document collection.

The framework can be applied to a visual sign dictionary as follows. First, requirements engineering is used to identify the information need/problem in RQ1. This does not only concern the Dutch translation of signs, but also other information which could support sign language learners. In RQ2 the matching process is addressed by defining sign similarity. The matching process is a key challenge, because perceived similarity between signs by sign language learners is not trivially defined. Next, RQ3 addresses the conceptual query and possible transformation. Users will express their search query by performing a sign in front of a webcam. The translation process will translate this search query to the system's document representation. Lastly, the conceptual analysis is addressed in RQ4 by exploring which aspects of a sign video are of most concern.

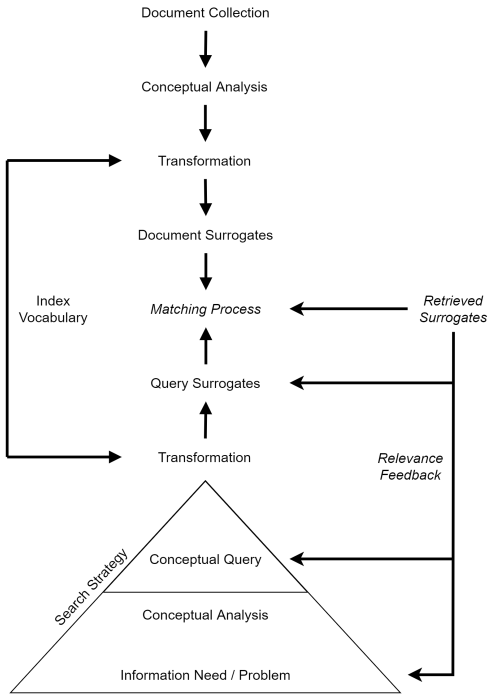

Figure 3.1: Information retrieval matching process between document collection and information need [40]

This IR framework supports this research by providing a clear overview of the connectivity between research questions and different parts of the visual dictionary. In addition, it gives a clear distinction between the user side and technological side of the system.

3.2 Motion Fused Frame

The document representation in the IR system is Motion Fused Frames (MFFs) introduced by Kopuklu et al.. A schematic representation of an MFF is shown in figure 3.2 [14]. This representation captures both spatial and temporal features by applying a data fusion strategy where RGB (Red, Green, Blue) images are combined with optical flow representations of preceding frames. These optical flow frames contain the estimated displacement field between two consecutive frames, showing the horizontal en vertical movement on a pixel level [45]. A video of a performed sign is a list of frames visualising the discrete states of the sign. An MFF contains a single state as RGB frame with movement information of preceding states as optical flow frames.

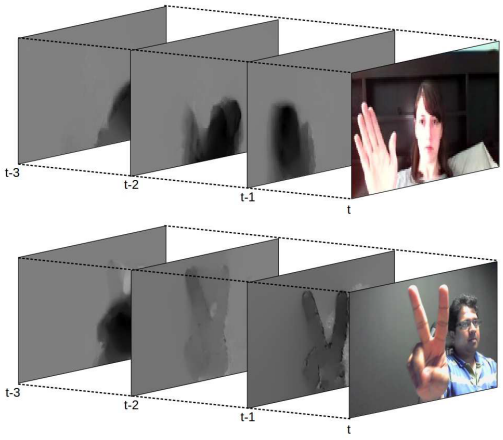

Figure 3.2: Motion Fused Frames architecture where optical flow frames are appended to RGB frames [14]

A sign is a series of states. An MFF captures both the current state in the RGB frames, the location and hand shape, as well as the movement between states with optical flow frames. This data format suits the sign structure by embodying both spatial and temporal features. Solely RGB frames would not capture the complete structure of a sign, since it excludes movement between states.

Flexibility in representation size is achieved by varying the amount of MFFs and optical flow frames. This data fusion method requires only one model to be trained, in contrast to decision or feature level fusion which requires one model for each data modality [7].

The MFF representation will function as a baseline when designing the transformation step in the IR process. Specifically, the MFF is the input to the neural network. Different MFF configurations are compared to the conventional RGB representations to determine the added value of the MFF representation and additional temporal information by varying the amount of appended optical flow frames.

3.3 Key Frame Extraction

The conceptual analysis determines what aspects of the documents are of most concern. Subsequently, the conceptual analysis is applied to the transformation process to select these aspects of the document which are of most concern. This means the most important frames need to be selected from videos of signs in the document collection and from recorded user queries in front of a webcam. These frames need to contain discriminative features to distinguish frames from different signs.

The current MFF algorithm uses uniform sampling. However, the discriminative properties of a sign are not linearly distributed over a sign. Each sign starts and end in the neutral position, frames containing these states would for example not be discriminative. Increasing the discriminative value of the selected frames will simplify model learning and potentially improve the relevance of the results.

Discriminative frames in a video of a sign can be defined according to the formational division of sign language morphemes created by Stokoe as shown in figure 3.3 [15]. This formational division is chosen because it shows the smallest building blocks of a sign which can change the meaning called phonemes. The introduced morpheme structure can be extended as a phonetic chart by listing all possible configurations. However, this research will extend on the morpheme structure by addition of the positions, configurations, actions, and locations that keep words of NGT separate to create a framework of NGT phonemes [15].

The phonemes described in the formational division can be described as orientation (forearm configuration), handshape (hand configuration), location (acted on) and movement (action). Signs

can also have nonmanual properties. These consist of mouthing where the spoken language equivalent of a word is articulated partly or as a whole without producing the actual sound and oral components where the articulation does not imitate an existing word [46]. The nonmanual component is not included in Stokoe's formational division.

These phonemes can be used to describe the discriminative value of a frame. This formational division is today seen as the proof for sign languages being an actual language, which emphasises its importance and validity [47]. The formational division will function as a framework to generically define the salient frames in a video of a performed sign by clarifying the structure and parameters of a sign.

This formational division was originally created for ASL(American Sign Language), however, both ASL and NGT fall under the same French sign language family [15]. The formational division is therefore likely to be applicable to NGT as well. Sign languages only differ in action and configuration, this formational division would therefore arguably be applicable to any sign language.

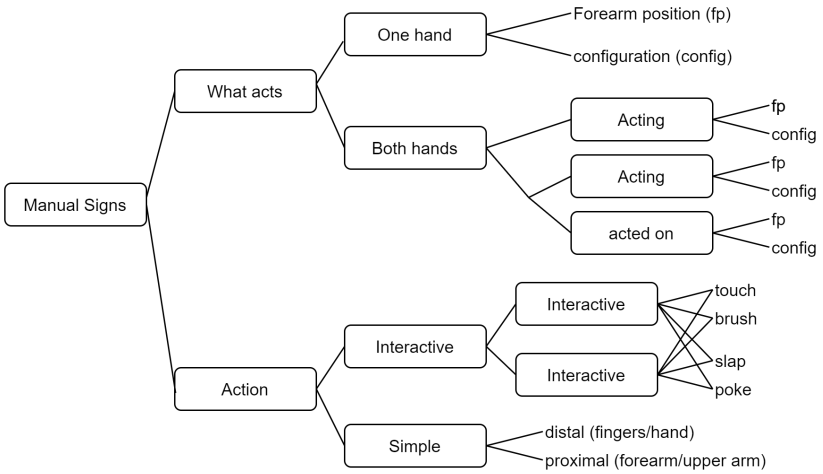

Figure 3.3: Stokoe's formational division of manual sign language morphemes

Chapter 4

Methodology

This chapter explains the research methodology. The first section discusses the research designs. Next, the data collection and analysis applied to answer the research questions are explained. Lastly, an evaluation of the research design and process is given.

4.1 Research Approach

The research followed a design and creation research approach that incorporates insights from interviews. The design and creation research explored the technical novelties concerning a visually searchable dictionary. Interviews were necessary to understand what this IR system should retrieve and how to evaluate the results. In addition, the interviews helped clarify which functionality sign language learners require from the application and how the user interface should look like. Thereby extending the scope from fundamental research to societal adaptation.

The interviews focused on how sign similarity is perceived by sign language learners. It can therefore be considered a qualitative research method with a phenomenological approach. The purpose of phenomenological research is to more fully understand the structure and meaning of human experience [48]. The human experience addressed in this research was the perceived similarity between signs by sign language learners. Whereas empirical research seeks to control and predict, phenomenology focuses on the process of understanding human behaviour from the individual perspective [49]. This phenomenological approach resulted in hypotheses, rather than testing hypothesis in empirical research. The interview was split in two parts which allowed to answer RQ1 and RQ2 separately.

4.2 Data Collection and Analysis

RQ1: Requirements Engineering

The interviews in this research are held with 7 people and are semi-structured. The interviews took place by video call and were recorded. The composition of interviewees is discussed in detail in section 5.1. The interview was composed using the book "Interviewen, theorie en training" (Interviewing, theory and training) to give confidence in the validity of the interview. The interview questions addressed the desired functionality and design of a visual dictionary that supports sign language learning and are listed in Appendix B. A motivation for the structure and content of the interview question is given in section 5.1.2. These questions were the second part of the interview covering both RQ1 and RQ2.

An inductive coding process was used for the analysis of the interviews [50] [51]. During the interview brief notes are made to streamline the analysis process. Afterwards, the recordings were viewed and in combination with the notes valuable quotes were written down for each participant for each question. Secondly, using open coding every statement was thematically clustered without

any filtering to explore the quotes, structure the quotes in a clear manner and identify themes [51]. The next step involves axial coding, where codes can be further split up, created and renamed [51]. This step will further structure the data by clustering the data within themes on lower level topic. The last step involves selective coding, where actual findings are deducted from the structured data and the research question is answered.

The selective coding process was adopted to RE by formulating the findings as user stories. This is a widely used format for formulating software requirements and states the user instance, desired functionality and reason for the functionality [34]. This analysis method should suit the research question by providing the requirements sign language learners have for a visual dictionary in a format that is suitable for software engineering.

4.2.1 RQ2: Sign Similarity

The second part of the interviews focused on which mistakes sign language learners make, what signs they find confusing and, subsequently, how they perceive similarity between signs. This concept was not well understood and interviews were used to construct in-depth insights about this topic. The interview questions are listed in Appendix A and a motivation for the content and structure of questions is given in section 5.1.1. This interview was held with the same 7 people as in RQ1 and was the first part of the interview session for RQ1 and RQ2. The same inductive process was applied, however the selective coding process consisted of formulating insights on how to define sign similarity. These insights together are applied to introduce and a quantification method for sign similarity as perceived by sign language learners. This sign similarity quantification was applied to an IR evaluation metric to quantify the relevance of results of a visual dictionary, thereby answering RQ2.

4.2.2 RQ3: Motion Fused Frames in Information Retrieval

RQ3 consist of a design and creation approach with an evaluation experiment. The design and creation research framework consists of five steps and is based on the book "researching information and computing systems" by J. Oates [52]. This method differs from conventional software development by addressing an unexplored problem, the design of a visually searchable IR system, without a commercial of-the-shelf software solution. During the design and creation research the problem is both theoretically analysed and evaluated in practice to form a complete understanding of the problem and solution.

This research method is structured as follows, the first step is awareness of a problem, which was the data representation in an sign language IR system. The suggestion step involved a proposed solution, substantiated by findings in literature. The development phase consisted of implementing the proposed solution. Next, the implementation was evaluated. Evaluation was performed on the NGT dataset, with a manually created test set. The evaluation metric defined and implemented in RQ2 is used to conduct these experiments. The results were performance scores for each MFF configuration. Analysis consisted of comparing the performance of these different MFF configurations. Lastly, a conclusion was drawn on how MFF compare to RGB as document representation in an IR system based on this analyses, thereby answering RQ3.

4.2.3 RQ4: OpenPose Key Frame Extraction

RQ4 applied the same design and creation approach as RQ4. The unexplored problem was key frame extraction in sign recordings. Next, the suggestion step consisted of conducting literature on key frame extraction, body keypoint detection and sign language structure to get a sound understanding of the unexplored problem. A solution was proposed and implemented. The evaluation was again performed with the same test set as in RQ3 and with the same evaluation metric defined in RQ2. Analysis consisted of comparing the evaluation scores for uniform sampling and the proposed key frame extraction method. Lastly, a conclusion was drawn on how key frame extraction using body keypoints extracted with OpenPose compares to uniform sampling, thereby answering RQ4.

4.3 Research Quality

4.3.1 Research design quality

The validity and reliability of this research have their limitations, which are effected by the quality of the research design. For RQ1 and RQ2 the number and variability of interviewees limits the completeness and generalisability of the results. Moreover, the results are based on NGT, application to other sign languages could be limited. Interviews fundamentally have their limitations as well, time constraints and question formulation may result in obtaining incomplete viewpoints of interviewees. The problems will be tackled by using an appropriate sample size and to ensure proper variability in the interviewees, interview design and question formulation. Sign language learners from various levels, sign language teachers, interpreters and researchers are approached to participate in the research. The interview design and interviewing techniques are based on the book "interviewing: theory, technique and training" by Ben Emans, to ensure the quality of the interview process [50]. Lastly data saturation will impact the validity of the interview results. Data saturation is reached when there is enough information to replicate the study, when the ability to obtain additional new information has been attained and when further coding is no longer feasible [53]. After analysing the interviews a judgement can be made about the saturation of the obtained data. Recurring statements by interviewees, absence of statements made by a single interviewee and absence of conflicting statements will be indications of data saturation.

RQ3 and RQ4 will consist of an experiment comparing the performance of different available and newly developed techniques. To ensure the reliability and validity only a single variable, the MFF configuration or key frame extraction method, will be changed. In addition, any form of randomness will be seeded, however, due to the parallel nature of the computation perfect reproducibility cannot be ensured.

The key frame extraction implementation in RQ3 needs to be evaluated with regard to the conceptual design. Due to low quality webcams, possible incorrect wrist detection of OpenPose and other factors the actual implementation of the conceptual design will not be a perfect implementation of the conceptual design. These imperfections need to be identified and listed.

Both RQ3 and RQ4 rely on a manually crafted test set to quantify the performance. This test set should imitate real life usage as closely as possible to make the results representable for real life performance. Any differences between the test set and real life usage could reduce the representability of the results in real life usage.

4.3.2 Research process quality

The implementation of the qualitative interview process used for RQ1 and RQ2 has a variety of limitations that impact the results. The interviews conducted for RQ1 and RQ2 are affected by, first of all, the sample size. With a sample fixed of of seven statistical validity is not guaranteed, but should give robust insights in individual perceptions. In addition, conducting the interviews via video calls could impact the results. Video calls are less personal and make it harder to read body language and to demonstrate examples. However, the ability to view the video recording did allow to view back the examples interviewees gave, which would not have been possible with a conventional audio recording.

Next, the students interviewed all studied at the same university. Sign language student from other universities could potentially have different views and provide other insights. Lastly, the interviews are not transcribed in their entirety, only valuable quotes are transcribed. This subjective process might result in quotes which are at first sight not considered valuable, but would have resulted in additional insights further in the inductive coding process. A larger sample size, conducting the interviews in person, interviewing sign language students from other universities and transcribing the entire interviews could have resulted in different or additional interview results.

The design and creation methodology approach applied to RQ3 and RQ4 is affected by the implementation process. A key decision was trimming the videos recorded for the test set to include only the neutral positions and signs, not the process of starting and ending the recording. These videos are trimmed because of the assumption that sign language learners would have to trim the

input videos themselves when using the visual dictionary. The processing of the input video is even discussed during the interviews. The interviewees unanimously considered the trimming of input videos to not be a problem, as long as the editing software is implemented in the visual dictionary. Unprocessed recordings would most likely have resulted in a lower performance.

Another design choice are the evaluation metrics used, which are $topKAcc$ and $NDCG@K$. Using different evaluation metrics to compare MFF to RGB frames and OpenPose key frame extraction to uniform sampling could have shifted the results for RQ3 and RQ4.

Chapter 5

Setup

This chapter discusses the setup of the interviews in RQ1 and RQ2, followed by the training strategy and model architecture used in RQ3 and RQ4. First the interview design is explained, followed by details on the interviewees. Next the design of interview questions for RQ1 and RQ2 is elaborated. Lastly, the model selection, training strategy and test set creation for RQ3 and RQ4 are discussed.

5.1 Interviews

The interviews for RQ1 and RQ2 are both designed with the book "interviewen: theorie, techniek en training" (interviewing: theory, technique and training) as framework [50]. In every step of the interview insights from this book are applied to improve the validity of the interview results. The interview is structured as follows.

The first step of the interview consists of a warm-up questions to introduce the topic and give the interviewee time to reminisce any memories regarding the topic. Before the interview started a few minutes are reserved to make the interviewee feel comfortable by introducing myself and asking some general questions to get to know each other. Next, the research is introduced as are the duration and privacy aspects. A fillable PDF consent form is sent before the interview. If the interviewee didn't fill in the consent form beforehand it was filled in at the start of the interview. During the interview attention is paid to respond neutrally to answers, both verbally and non-verbally. In addition, interview questions are carefully constructed to have a neutral tone. Attention is paid to keep any follow up questions neutral as well to prevent any implication of answers. These follow up questions aim to make interviewees think more carefully about a certain topic to get a more comprehensive answer. Whenever an interviewee is having trouble with constructing an answer a reminder was given that there were no time constraints. If the interviewee was still having trouble constructing an answer hints were given on generic topics to cover in the answer. Each interview part is finalised by asking the interviewee if there are any other things that came to mind in general about the topic. This measure should get the interviewee freely talking about the topic to discuss anything that wasn't covered by a specific question. These measures should ensure the validity of the data generated during these interviews.

Interview participants consisted of a sign language teacher, PhD student on sign language classification, sign language interpreter and five sign language students of varying study progress. The PhD candidate did not speak NGT, but international sign language (ISL) and American sign language (ASL). His extensive knowledge on sign languages should however give valuable insights on sign similarity, which should be generalisable to NGT. The NGT interpreter and teacher work for the same organisation and were interviewed simultaneously. The NGT teacher is a native NGT user and deaf from birth, the organisation provided an interpreter to communicate during the interview. Lastly, the NGT students have varying study progress, varying from a semester long minor to a fourth year student. The interviewees backgrounds are listed in table 5.1.

Interviewee	Sign language background	Sign Languages
1A	NGT Interpreter	NGT, ASL
1B	NGT teacher	NGT
2	4th year student NGT interpreter	NGT
3	PhD on sign language	ISL, ASL
4	2nd year student NGT teacher	NGT
5	30 ECT minor sign language	NGT
6	followed three courses and thesis on NGT	NGT

Table 5.1: Interview participants

5.1.1 Sign Similarity

The central question to be answered during this interview part is how sign language learners perceive similarity between signs to implement a validation metric for measuring the relevance of the retrieved results by the visual dictionary. The interview design had a funneling approach, starting with broad questions on confusion signs and mistakes during signing and thereby implicitly on sign similarity. Next was a rather direct question on how to define sign similarity, which was used to test what the first thing was that came to mind when thinking about sign similarity. Stokoe's taxonomy was then introduced where confusability caused by position, hand/arm configuration and movement were discussed in isolation. The applicability of the taxonomy to define sign similarity was discussed next, followed by a general question on what users wanted to see back in the results. The interview is finalised with an open question to discuss anything that comes up regarding sign similarity.

5.1.2 Requirements Engineering

The second part of the interviews concerned the requirements potential users have for the system. Questions mainly concerned additional functionality next to retrieving videos of sign to support sign language learning. The questions regarding functionality should result in user stories which software engineers can use to implement the desired functionality of sign language learners. Other questions concerned the usage process and user interface to clarify constraints and requirements. These questions combined should result in a clear and complete set of requirements and constraints in the form of user stories. These user stories should ensure that the functionality of the developed visual dictionary meets the desired functionality of its users.

5.2 Model Selection

With the advent of deep learning in recent years the research field on sign recognition has developed greatly [22]. Many different implementations are available, which use a variety of data modalities, such as RGB (Red, Green, Blue), depth, skeleton and thermal. Of this wide variety of models only a selection is open source, of which the data level fusion strategy called Motion Fused Frames (MFF's) is been selected for its single model approach. A decision or feature level fusion strategy would require multiple models to be trained, increasing the complexity of model design training and inference.

MFF's consist of RGB images concatenated to optical flow frames. These MFF's can be configured with varying number of MFF's N , optical flow frames x and color frames y, denoted by N-MFFs-xfyc. Increasing N results in more spatial information, whereas increasing f results in more temporal information. This video representation can transform a video of arbitrary length to a fixed number of frames which is typically required for machine learning models [7][6][14]. This video representation has been successfully applied to the Jester dataset [14]. This dataset contains 148,092 videos of gestures performed by 1300 volunteers recorded in front of a webcam in an unconstrained

environment [54]. Both the dataset format, recording setting and gestures, are similar to the visual dictionary, where users can search for signs by performing an NGT sign in front of a webcam in an unconstrained environment. In addition, this data fusion format requires only a single model to be trained, reducing implementation and training complexity. Lastly, the source code was available on GitHub with installation documentation [55].

The model architecture uses a convolutional neural network (CNN) as backbone to create an image embedding out of the MFF's. The CNN backbone used in the paper is BN-Inception, which achieves a top-1 accuracy of 74.8% on the ImageNet classification task [56]. This CNN has been replaced with the EfficientV2-S model, which achieves a top-1 accuracy of 84.9% [57]. The CNN architecture is created using a neural architecture search which is set to optimise accuracy, parameter efficiency and training efficiency. Both models are compared in table 5.2. This better performing CNN backbone could results in increased accuracy and training time performance compared to the original CNN architecture used. The next section will discuss the training process of this model on the NGT dataset.

Model	**Params**	**GFLOPS**	**Val Top-1 Acc.**
BN-Inception	11M	2.2	74.8
EfficientNetV2-S	22M	8.8	84.9

Table 5.2: Comparison between the original CNN used as backbone in the paper and proposed CNN [14]

5.3 Transfer Learning

One-shot-learning (OSL) approaches without appropriate regularisation generally results in overfitting. This means a model with Hypothesis space H, training set D_{train} and test set D_{test} is trained to a hypothesis h_{train}, which does not generalise well to D_{test} [41]. The optimal hypothesis $\hat{h}$ is generally far from h_{train} with a small training set, which is denoted with h_I. To tackle this problem an algorithmic and data driven approach are applied. The algorithmic approach consists of several transfer learning steps, where the model is initialised with parameters θ_o, which are expected to be closer to $\hat{h}$ than h_r with random initialised parameters θ_o. The transfer learning step is thus expected to result in parameters which are closer to the optimal parameters than randomly initialised parameters.

This process is repeated to fine tune the parameters with a training set which is more similar to the actual training set. The transfer learning process applied before training on the NGT dataset D_{NGT} is shown in figure 5.1. The random initialised parameters are first fitted to the huge ImageNet21K training set to learn generic features. Secondly, the model is fine tuned on the large ImageNet1K dataset. Whilst ImageNet21K and ImageNet1K are extensive datasets, they contain everyday images of for example animals and vegetables, which are conceptually far from the sign language frames in the target dataset. Moreover, these models are trained on RGB images, leaving room for improved applicability on MFFs.

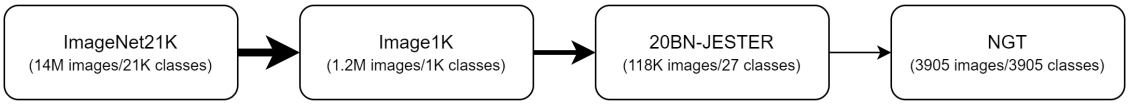

Figure 5.1: Transfer learning process from the huge general ImageNet21K dataset to one-shot learning NGT dataset. The first two steps generic features are learned, whereas the third step learns features close to the sign language domain.

The Jester dataset containing gesture video's is used to fine tune the parameters to a domain closer to the target domain and target data format as a third transfer learning step. To artificially increase the size of the training set a variety of data augmentation techniques are applied to increase I to $\tilde{I}|\tilde{I} \gg I$ to obtain a more accurate empirical risk minimiser $h_{\tilde{I}}$, which is closer to h^* than h_I [41].

These data augmentation techniques should ensure a richer training dataset, reducing overfitting. Transfer learning and data augmentation should together enhance the one-shot learning capabilities.

5.3.1 NGT Validation and Test Set

By definition, a one-shot learning problem does not have any train/test split possibilities, since there is just one sample per class. Additionally, the provided training videos are recorded under studio conditions and the signs are articulated by a experienced signer. This results in videos of high quality, no background noise and flawless articulation. In practice, however, the recordings will be made with a webcam in home environments. The signs will likely not be articulated flawlessly, since sign language learners can make articulation errors or will look up signs which they do not remember correctly. The training sample are thus not representative for the use case of the system. Even when multiple samples per sign would be provided, the validation and test metrics would not correctly reflect the performance of the IR system in practice.

Figure 5.2: Provided studio frame and self recorded frame of the sign for zalf (ointment)

To get a representative validation and test set six volunteers, including me, have recorded 253 videos of signs in home environments using the build in laptop webcam or external webcam. Five out of six volunteers have no signing experience and one is an experienced signer. Instructions were given to view each sign and practice its articulation before recording the sign. This is deemed necessary to ensure a minimal articulation level by participants with no signing experience. In addition, the recording environment is not cleaned up and the recording location is the usual computer usage location. Participants generally used a conventional desk as their recording location. Two exception were a participant using a standing desk and a participant using a couch. In some cases the office chair was required to be located slightly more backwards than usual to get the whole signing space in the recording. This was the only adaptation made in comparison to the natural computer usage setting of the participants. Moving the office chair slightly backwards to allow the webcam to record the whole signing space will be required for adequate performance, as otherwise signs could be performed out of view. The instructions, home environments and conventional webcams should result in validation and test videos which closely reflect usage in practice. The 253 recordings are split in 100 validation samples and 153 test samples, stratified with respect to the participants. An overview of the participants and recording details can be found in table 5.3.

Participant	**#Val Rec.**	**#Test Rec.**	**Total**	**Recording Location**
Participant 1 (location 1)	23	33	56	Couch
Participant 1 (location 2)	20	30	50	Desk
Participant 2	15	23	38	Desk
Participant 3	12	18	30	Standing Desk
Participant 4	11	18	29	Desk
Participant 5	11	16	27	Desk
Participant 6	10	16	26	Desk
SUM	100	153	253	

Table 5.3: Number of validation and test recordings per participant and the recording location.

Chapter 6

Requirements Engineering

This chapter discusses the results of the requirements engineering part of the interviews. The first section elaborates the inductive coding process and the second section discusses the results from the coding process. These requirement should provide a comprehensive overview of the desired functionality and design from sign language learners of a visual dictionary.

6.1 Inductive Coding

The second part of each interview consisted of question regarding the requirements sign language learners have from a visually searchable sign IR system. The interviews were analysed using an inductive coding process of which the result is shown in table 6.1 [51]. The first step was to identify open codes for broad topics in the quotes, these are shown as headers with grey background. The themes identified were "use case", "result layout", "additional information" and "video processing". Next, the axial codes were identified, where the structured quotes are grouped in specific topics. These topics are listed in the left column under the open codes. Lastly, the selective codes are determined, which are the insights obtained from the quotes in the axial code formulated as user stories. These user stories provide answers to the original question on which desired functionality and design sign language learners have from a visual dictionary. The selective codes are listed in the right column within a theme. The next section will extensively discuss the obtained insights.

Use Case	
Axial Code	**Selective Code (User Story)**
Primary Use Case	As a user I want to primarily use the tool as a dictionary, where the IR use case is of secondary importance, such that I can look up a specific sign
Information Retrieval Design	As a user I want to retrieve multiple results, such that I can find a sign when I do not sign properly
Semantically similar Results	As a user I want to retrieve semantically similar signs, such that I can learn signs related to the context of the sign I am looking for
Instructions to Users	As a user I want to get concise instructions on how to record the sign such that I correctly use the tool
Result Layout	
Axial Code	**Selective Code (User Story)**
Number of Results	As a user I want to view retrieve around 10 results, such that the results are balanced between conveniency and completeness
Increase Number of Results	As a user, I want to be able to increase the number of results to view more similar signs, such that the initial results are clear, but I can keep searching if I can not find the sign directly
Confidence Dependent Results	As a user I want to get more or less results depending on how confident the retrieval system is, such that I do not get unnecessary results
Clean User Interface	As a user, I want to have a minimal user interface, such that the tool is easy to use
Additional Information	
Axial Code	**Selective Code (User Story)**
Iconicity of Sign	As a user, I want to see the origin/iconicity explanation of a sign, if there is one, such that I remember signs better
Usage of Sign	As a user, I want to see in which context a word can be used and in which context not, such that I learn to correctly use the sign
Signs From Side View	As a user, I want to be able to see a sign from the side view, such that I can get a better view of movements in the z-axis
Video Processing	
Axial Code	**Selective Code (User Story)**
Trim Video in Tool	As a user, I want to trim the video in the tool, such that I do not need to use a second program to use the tool
Video Editing Explanation	As a user, I want to see an explanation why the video needs to be edited, such that I understand why this extra step is needed
Security	As a user, I do not want my video to be viewed by other users, such that my privacy is ensures

Table 6.1: Inductive coding analysis of requirements engineering interviews. For each open code the axial and selective code are listed, where the selective is formulated as a user story.

6.2 Insights

The user stories give a comprehensive overview of the requirements users have from a visually searchable sign IR system. The first identified open code concerns the use case of the visual dictionary. Users would, first of all, primarily use the system as a dictionary, not as a tool to look up similar signs. Although looking for a single result, users anticipate for sign language learners to not articulate properly, requiring multiple results to find a sign. When users find a sign they would like to have the ability to see semantically similar signs, as this would help them to extend their vocabulary in the context they are working in.

The next open code concerned the result layout. Interviewees mentioned varying desired number of results from 6 to 20, where 10 seemed an appropriate balance between conveniency and completeness. Users requested functionality to increase the number of results when desired and hinted on a mechanism where the number of results was dependent on the level of confidence of the IR system. This would show experienced signers with a high quality webcam fewer results than a inexperienced signer with a low quality webcam. Interviewees also desired a minimal user interface to make the tool intuitively to use.

Additional information on signs is covered in the third open code. Desired additional information consists of iconicity explanation. Linguistic iconicity is defined as the existence of a structure-preserving mapping between mental models of linguistic form and meaning [3]. Iconicity explanation would thus elaborate on what the sign should resemble. For example, the NGT sign for "thee" (tea) resembles the dipping of a tea bag in a mug. This would make a sign more intuitive and therefore easier to remember. Moreover, certain Dutch translation of signs are homonyms which have multiple signs in NGT. An example would be the Dutch word "bos", which could either mean forest or bunch. These two homonyms have different signs in NGT. This inconsistent mapping between Dutch and NGT is a source of mistakes. Sign language learners mentioned information on the correct and incorrect usage contexts of a sign would help them correctly apply a sign. Other additional information would include sign parameters, such as hand shape and location, as these are sometimes hard to correctly interpret from a video. Next, interviewees mentioned some signs have movement in the z-axis, for which it is hard to interpret the sign from a frontal view. Having functionality to view signs from a side view would, for these signs, give a more clear view of the sign.

The last open code addressed the video processing of input videos. Recorded videos of users performing a sign need to be trimmed before feeding to the IR system, such that the video contains only the articulation of the sign. Interviewees mentioned they would prefer to use a video trimmer build into the visual dictionary, instead of trimming the video using their own video editing tools. The process would also require en explanation for the goal of the trimming, namely improving results. A non-surprising requirement was the privacy of user, they would after all upload videos of themselves to the system. Other users should not be able to view the videos of other users.

6.3 RQ 1 Conclusion

During the interviews a variety of requirements are elicited from sign language learners, which are potential users of a visually searchable sign dictionary. These requirements include additional information on retrieved signs, usage process and user interface requirements, as well as security requirements. These requirements are formulated as user stories and should guide developers to create a visually searchable sign IR system that suits the users needs.

Chapter 7

Sign Similarity

This chapter covers the process of defining perceived similarity between signs by sign language learners and the corresponding implementation of a similarity score. First, the inductive coding process is explained, followed by a similarity quantification per phoneme. Next, the complete sign similarity quantification formula is presented and lastly RQ2 is concluded. The sign similarity quantification constructed in this chapter supports the evaluation metric for the visually searchable sign IR system. In addition, the interviews give insights in the learning process of sign language students and the way they distinguish and confuse signs.

7.1 Inductive Coding

The interviews were held with seven participants divided over six interview sessions. The interviews are analysed using an inductive coding methodology [51]. This resulted in the code tree shown in figure 7.1. Four specific open codes were induced from the quotes, namely location, nonmanual, movement and handshape. A fifth code was assigned to all quotes without a clear theme. These open codes denote the broad themes identified in the quotes during the first analysis round and are shown in the right rectangles in figure 7.1. The four open codes correspond to sign language phonemes. Axial codes were assigned to clusters of quotes covering the same topic to further group the quotes. The axial codes are induced by grouping quotes within a open code that cover the same specific topic. Axial codes are shown on the right in figure 7.1 and denote the specific topics covered within the theme of an open code.

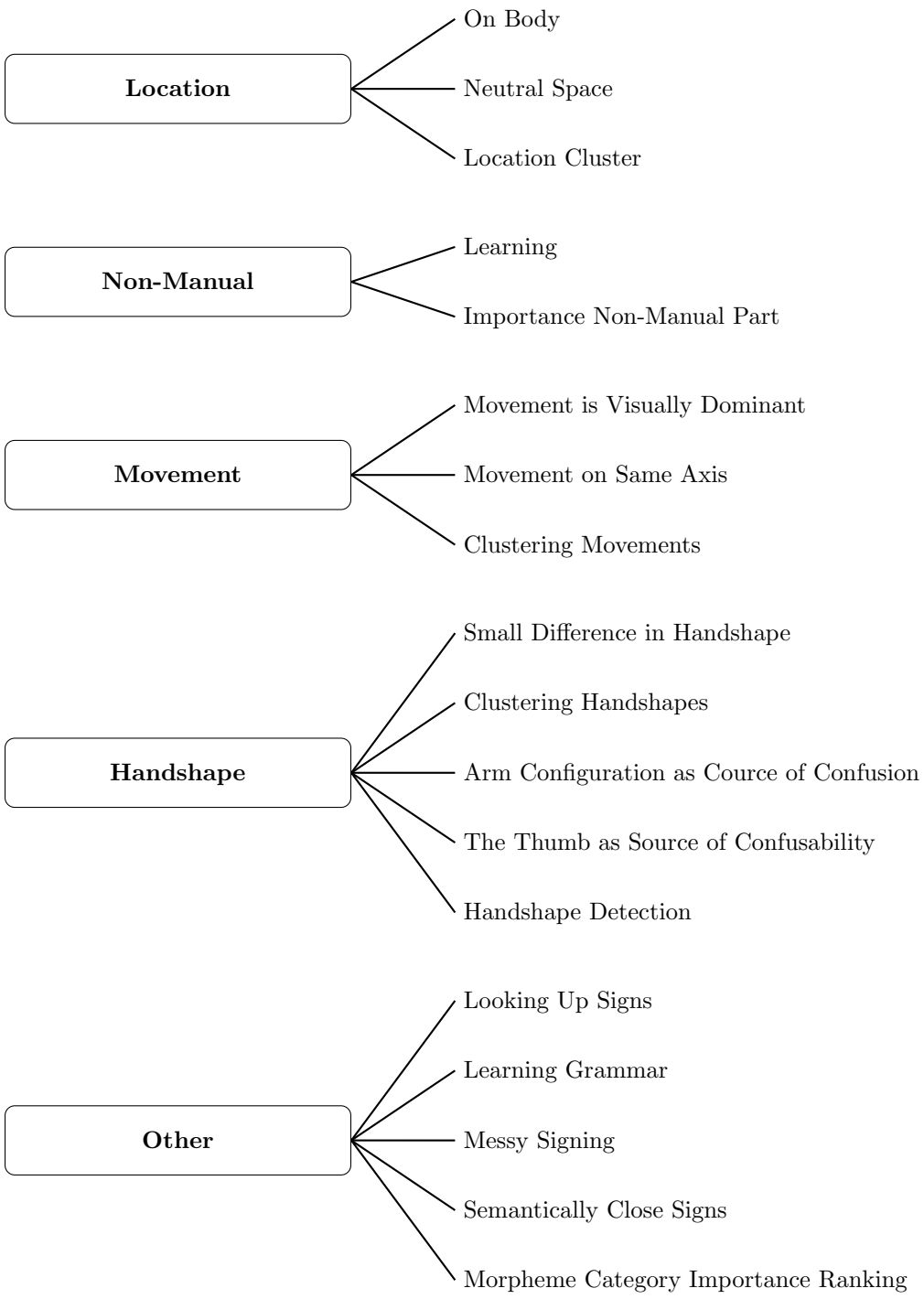

Figure 7.1: Open and axial inductive codes of the interview quotes. Open codes denote an overarching theme and axial codes further specify a group of quotes within a theme.

The last step consisted of assigning selective codes, these were constructed as sentences given the core insight of a given cluster of quotes within an axial code. These axial codes denote the generic message from a group of quotes covering the same topic. The insights should be of value with respect to the research question: how to define sign similarity. In table 7.1 the open, axial and selective codes are shown. Open codes are denoted as grey headings, axial codes are listed in the

left column within an open code and selective codes are listed in the right column within an open code.

Position	
On Body	Different positions on the body are confusing
Neutral Space	There is variety in the neutral position, this is however not a source of confusability and should not be differentiated
Location Clusters	positions can be clustered as follows: neutral space, lower and upper torso, weak hand and especially the upper and lower head
Non-Manual	
Learning	Sign language learning tend to focus first on the visual manual aspect and later on the nonmanual part
Importance nonmanual part	Non-manual part is important for understandability, however sign language learners tend to focus on the manual aspects and wouldn't search on nonmanual aspects
Movement	
Movement visual dominance	Movement is visually dominant
Movement on same axis	Movements on the same axis are visually similar
Clustering Movements	Movements should be clustered on axis
Handshape	
Small Difference in handshape	Small differences in handshape cause confusion, many handshapes are visually similar
Clustering Handshapes	Handshapes can be clusterd, but clusters will be relatively small
Arm Configuration as source of confusion	Arm configuration is not a major contributor to sign confusability
Thumb as source of confusion	Thumb configuration is a source of confusability
Handshape Detection	Handshape is expected to be hard to detect, especially compared to movement
Handedness	Handedness is an efficient way to downsize the results
Other	
Looking Up Signs	Sign language learners have difficulties with searching for signs through sign parameters
Learning Grammar	Grammar is difficult when learning sign language
Messy Signing	Messy signing can result in variability and inprecise articulation
Semantically Close Signs	Semantically close words can be confused
Morpheme Category Importance Ranking	Movement is an important indicator for sign similarity and is visually dominant, position is also easy to remember, but handshape is hard to detect and easily mistaken

Table 7.1: Open, Axial and Selective inductive codes of the interview on sign language similarity. Quotes are firstly clustered in open codes which are denoted in grey headings an cover a general theme. Within an open code, quotes are clustered on a specific topic where the selective codes cover the central insight induced from these quotes.

7.2 Information Retrieval Relevance Scores

Selective codes answer the research question with conventional inductive coding. However, in this research these insights form the basis for implementing an evaluation metric for the IR system. The following four subsections will discuss how the selective codes are used to construct a theoretical framework for sign similarity on four phonemes: location, movement, handshape and handedness. Lastly, the four similarity scores are combined to introduce a quantification method for sign similarity in section 7.2.1.

Non-manual aspects are left out of scope, because the provided dataset did not contain proper nonmanual labels. The entries "Mouth gesture" and "Mouthing" have been added to each label, however just 19 out of 4157 signs had a mouth gesture label, whereas mouthing labels were not provided at all. With no other nonmanual morphological labels provided the dataset's nonmanual information richness was considered insufficient to construct a nonmanual distance metric. Arm configuration was also left out of scope, because only the configuration change was provided, not the static configuration. Additionally, only 1733 out of 4157 orientation change entries were provided, of which 1196 were left blank, resulting in 537 actual values. Arm configuration was also not seem as a major source of confusion by the interviewees. This all resulted in the exclusion of arm configuration of the sign similarity metric.

7.2.1 Location

During the interviews participants mentioned locations close to each other were confusing. This closeness should, however, be unambiguously defined. A distinction could be made between locations on the body and locations in the neutral space, the area in front of the torso. The neutral space allows for variety in locations of signs. Participants indicated this variability is not a source of confusion. The assigned location labels do not subdivide the neutral space. This aligns with the viewpoint of participants that the neutral space should be seen as an atomic location where variation does not cause confusion.

Different locations on the body are, in contrast to the neutral space, viewed as a source of confusion. By combining statements from interviewees regions could be identified in which different locations could cause confusion. Firstly, different location on the head are viewed as a major source of confusion. Locations on the head were such fine grained they had to be further subdivided into the upper and lower head. Secondly, the lower, and especially, the upper torso were considered regions. Lastly, the weak hand was viewed as a separate regions of locations. All location labels in the dataset could be mapped to a regions identified by interviewees, except locations on the arm and beneath the torso. These locations were assigned to two new regions, respectively arm and sub torso. The locations were structured using a taxonomy tree, a visualisation is shown in figure 7.2. Leaves are listed for readability.

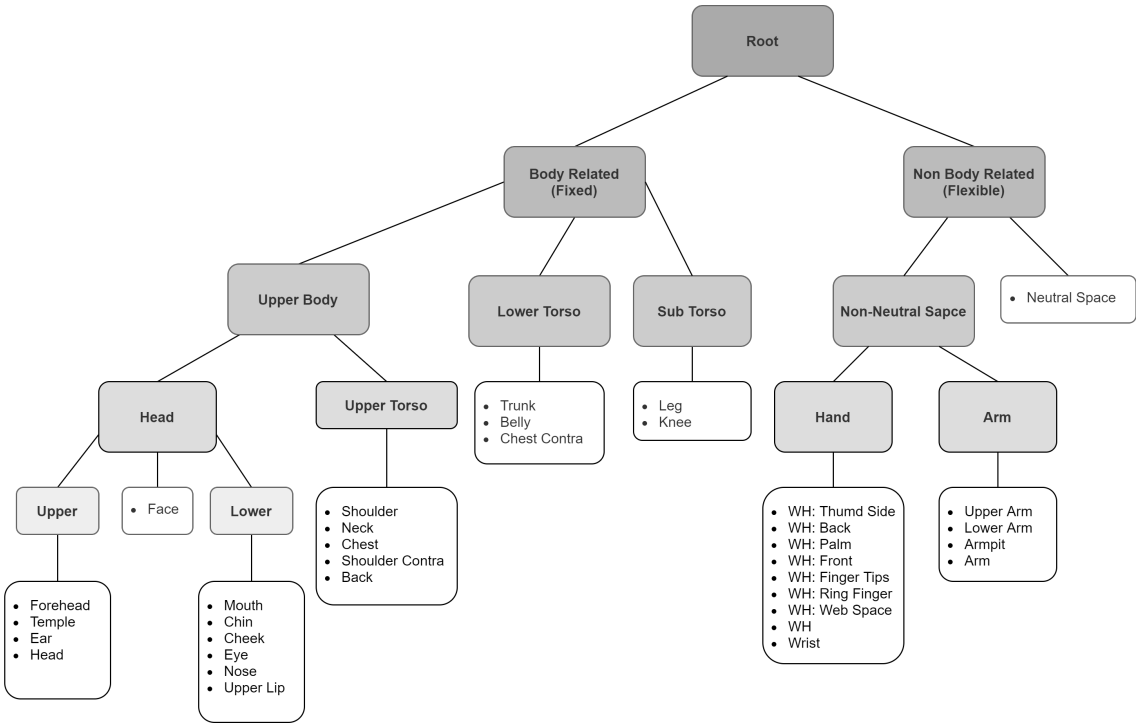

Figure 7.2: Sign location taxonomy tree where leaves are itemised for readability purposes. WH stand for Weak Hand. The locations within a region are easily confused with each other.

Similarity quantification is computed by subtracting one by the node distance between leaves as fraction of the longest distance, resulting in a similarity score in the practical range $[0, 1]$. The node distance is the amount of nodes required to travel from one location to another location, which is equal to the amount of edges between two locations. The node distance is subtracted from one because a longer node distance should result in a lower similarity score.

The longest node distance is nine, which is between upper/lower head and hand/arm. For example, the similarity between the locations forehead and shoulder is $1 - \frac{5}{9} \approx 0.44$. The distance is the path $forehead \Rightarrow upper \Rightarrow head \Rightarrow upperbody \Rightarrow uppertorso \Rightarrow shoulder$, which are five steps. Similarity between forehead and ear is $1 - \frac{2}{9} \approx 0.78$. The listed leaves should thus be interpreted as single leaves, making the path between forehead and ear $forehead \Rightarrow upper \Rightarrow ear$.

Movement

Discussing similarity in movement participants unanimously pointed to movements articulated on the same axis as similarly looking signs. The direction was not seen as a discriminative feature. Signs performed from left to right could therefore be confused with signs performed from right to left, as long as they are performed on the same axis. Participants also mentioned the movement being visually dominant, making it a key phoneme to validate the relevance of the results with. If all signs retrieved have similar movement to the input sign the results would have a high relevance for the user.

The movement direction labels provided in the dataset contained axial labels to cluster on. Movements are clustered in movements in the vertical and horizontal planes. Horizontal movements are subdivided in the x and z axis. Movement consisting of a vertical and z-axis component were assigned under diagonal. Signs with no movement were clustered under "None". Lastly, several special movement labels, such as "variable" and "from location", were grouped under "other", because no clear nor consistent movement axis could be determined. The z-index, consisting of forward and backward movement, were deemed hard to interpret from a frontal view. The z-index was therefore ignored when combined with another axis. For example, the movement "backwards

and downwards" is assigned under vertical movement. As with locations, the movement clusters are structured in a taxonomy tree to visualise the distance, this tree is shown in figure 7.3

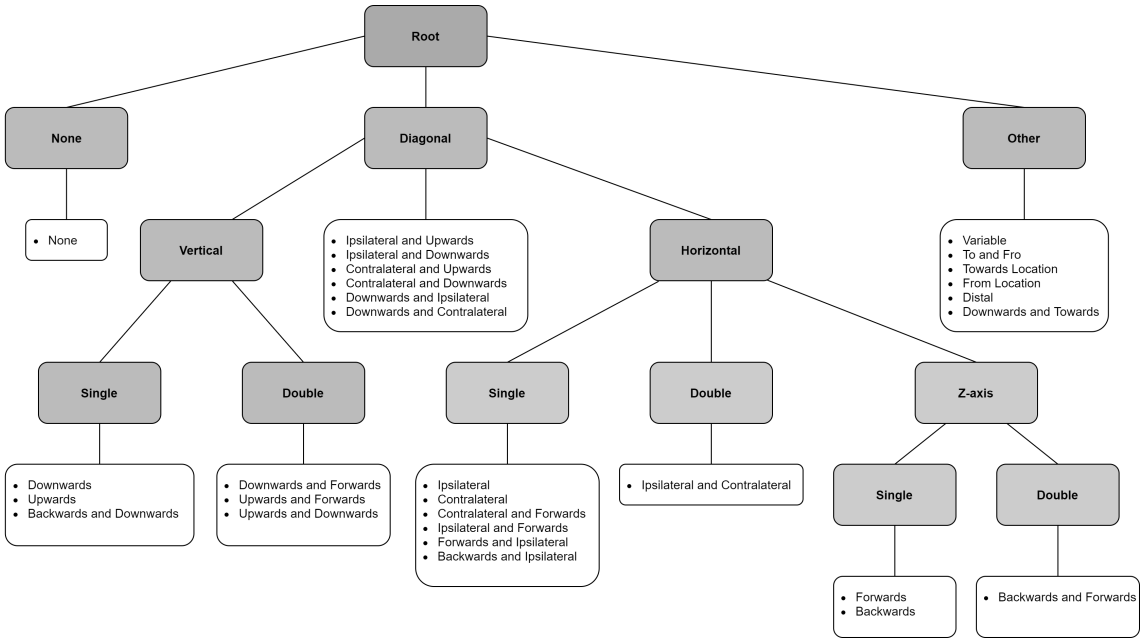

Figure 7.3: Sign movement taxonomy tree where leaves are itemised for readability purposes

Several design choices have been made to best structure the distances as shown in figure 7.3. Firstly, diagonal movement is placed between vertical and horizontal movement, because diagonal movement is considered visually in between vertical and horizontal movement. This makes vertical and horizontal movements equally similar to diagonal movements, but more dissimilar to each other than to diagonal movements. Secondly, movements on the z-axis are considered visually similar to other movements on the horizontal plane, but dissimilar to diagonal and especially vertical movement. Therefore, movements on the z-axis are assigned under horizontal movements, making them similar to x-axis movements, but dissimilar from diagonal and especially vertical movements.

Just as with locations, similarity is quantified as the node distance between movements in proportion to the maximum node distance in the tree. The maximum distance is seven, which is for example the distance between forwards and downwards. The similarity between forwards and ipsilateral (left to right) would for example be $1 - \frac{|single, z-axis, horizontal, single, ipsilateral|}{7} = \frac{5}{7} \approx 0.29$.

This structure can convert any pair of movement directions to a similarity value in the range $[0, 1]$, based on visual axial similarity.

Handshape

When discussing similarity in handshape during the interviews, participants mentioned there were many similar looking signs, such as the B/B-null, C/O and counting hands. Generic properties between different handshapes that make signs visually similar involve the amount of fingers used in a sign, thumb configuration and curve of the fingers. The labels provided in the dataset did however only mention the handshape, not the properties of the handshape.

The PhD thesis of van der Kooij on phonological categories of sign language in sign language of the Netherlands provided a mapping from handshape to articulator properties [3]. These articulator properties assign generic properties to handshape used to define handshape similarity, which should reflect the generic handshape properties interviewees had difficulty naming. For example, the B and B-null handshape, which are considered visually similar by one of the interviewees, both map to the categories "all". These handshape are thus considered visually equal according to the handshape

properties too. Signs can also map to multiple articulator properties. For example the C and O handshapes both map to "all", "open" and "curve", denoting all fingers are extended, the thumb opposed to but not touching the selected fingers and flexion of at least the non-base joints.

Using these articulator properties a similarity quantification has been constructed based on the fraction of the intersection of articulator properties with respect to the union of articulator properties. This similarity quantification method is chosen to capture the amount of overlapping handshape properties. A higher fraction of overlapping handshape properties is assumed to result in higher visual similarity.

This results, again, in a similarity score in the range $[0, 1]$. The similarity between the B and C handshape would for example be $\frac{|\{all\}|}{|\{all,open,curve\}|} = \frac{1}{3} \approx 0.33$. When all properties are equal the similarity score is 1 and when none of the properties are equal the score is 0. Similarity scores are given for the strong and weak hand separately. In asymmetrical signs the strong hand is the moving hand, whereas the weak hand serves as a location for the strong hand [3]. Symmetrical signs do not have a strong or weak hand separation, since signs both hand are moving symmetrically and both hands have the same handshape. A similarity score for both the weak and strong hand results in two similarity scores for hand shape.

The similarity quantification method does treat every property similar and does not take into account the additional specifications provided for simplicity reasons. This method should however form a baseline to systematically define similarity in handshapes based on generic properties.

Handedness

There were no specific questions on handedness, however one interviewee did mention handedness was an efficient way to filter down the signs, indicating its importance for the relevance of the results. Handedness is taken into account for quantifying sign similarity as it is considered visually dominant by interviewees, even stating confusion one and two handed signs to be extremely rare. In addition, handedness is a phoneme in Stokoe's formational division of sign language morphemes [15]. Moreover, Beluggi's research on remembering signs did mention the handedness was preserved when signs need to be reproduced and errors were made in other phonemes. Handedness is thus assumed to be a source of similarity based on previous research. Covering the role of handedness in sign similarity could have further substantiated its importance.

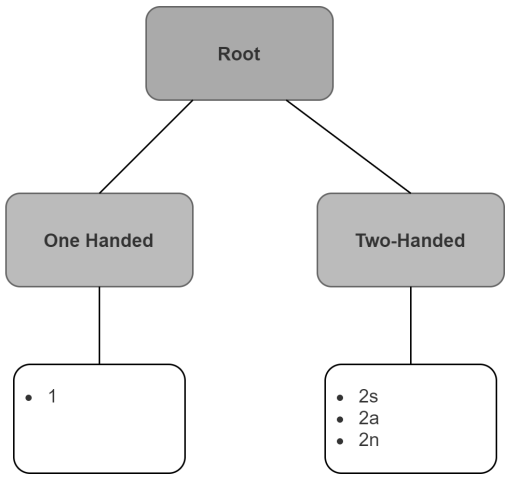

Figure 7.4: Handedness taxonomy tree where leaves are itemised for readability purposes

The handedness similarity is, firstly, divided into one-handed and two-handed signs. The latter is subdivided into different configuration. The result is the taxonomy tree shown in figure 7.4. Distance between handedness configuration is measured as node distance proportional to the maximum distance, as with location and movement. The maximum distance is four, which is between a one-handed and a two-handed configuration. The distance between different two-handed configurations is $1 - \frac{2}{4} = 0.50$.

Sign Similarity Formula

The four phonemes location, movement, handshape and handedness assembled quantify sign similarity. Handshape is subdivided into strong hand and weak hand, resulting in five similarity score. The final similarity score is the mean of these five similarity scores, resulting in a sign similarity quantification in the range $[0, 1]$.

This sign similarity score is applied as relevance score in the Normalised Discounted Cumulative Gain (NDCG) metric to measure the relevance of the retrieved signs [13]. The NDCG formula is shown in equation (7.1). Here, Q denotes the number of queries, iN_i the maximum possible $DCG@K$, d^i_j the j_{th}-ranked sign returned by the model in response to query q_i and $rel(d^ij)$ the relevance of the j_{th} retrieved document with respect to query q_i.

$$NDCG@K = \frac{1}{Q}\sum_{i=1}^{Q}\left[\frac{1}{N_i}\sum_{j=1}^{K}\frac{2^{rel(d^ij)}-1}{log(j+1)}\right] \tag{7.1}$$

The relevance of a document is given by $\frac{2^{rel(d^ij)}-1}{log(j+1)}$, where lower signs are counted less towards the score with division by $log(j+1)$. The $rel(d^i_j)$ relevance of a sign for a given input sign is quantified using the introduced sign similarity quantification. Eventually the score is divided by the maximum DCG score to get a value in the range $[0, 1]$. The NDCG score is thus the relevance of all retrieved signs proportionate to the maximum achievable relevance of the results.

7.3 RQ2 Conclusion

Interviews with a variety of stakeholders from the deaf community identified four phonemes which together define sign similarity. These phonemes are location, movement, handshape and handedness, where handshape is subdivided into weak hand and strong hand. During interviews generic properties to define sign similarity are identified for location and movement. Handedness properties are based on literature.

For location, movement and handedness a taxonomy tree is constructed where similarity is defined as node distance with respect to the maximum node distance in the tree. Location similarity is divided into regions in which sign language learners easily confuse locations. These regions are upper and lower head, upper, lower and sub torso, neutral space, weak hand and arm. Movement is characterised by the axis it operates on, where the direction is not perceived as a discriminative property, movement is therefore divided by axis. Handedness is simply divided in one and two handed signs, where two handed signs are subdivided in symmetrical and asymmetrical signs. Generic properties for handshapes could not be determined during interviews and the articulator properties based on the work of van der Kooij are used instead [3]. Handshape similarity is defined as the fraction of the intersection of articulator properties of two handshapes with respect to the union of their articulator properties. A separate similarity score for both the strong and weak hand are given.

The mean of the five separate similarity scores is the final similarity score between two signs in the range of $[0, 1]$. This similarity score is is used for evaluation in RQ3 and RQ4. Besides the derived similarity score, these findings also give a comprehensive analysis on how sign language learners perceive similarity between signs.

Chapter 8

Motion Fused Frames in Information Retrieval

This chapter describes the results of the training process and evaluation of the visually searchable sign IR system. The first section lists the results of the Jester transfer learning process, followed by an first experiment to evaluate the model on the NGT dataset without a test set. Next the hyper parameter optimisation process on the NGT dataset with an evaluation set is described and the evaluation using the test set. Lastly, the performance is quantified using the Normalised Discounted Cumulative Gain (NDCG) metric where the sign similarity quantification constructed in chapter 7 is used as relevance score. The central topic in this chapter is how the Motion Fused Frames (MFFs) data representation perform in an IR setting compared to Red, Green and Blue (RGB) frames and how this can be appropriately evaluated.

8.1 Jester Transfer Learning

The original MFF model architecture described in the paper by Kopuklu et al. is modified by replacing the Convolutional Neural Network (CNN) backbone and expanding the transfer learning process with an extra step on ImageNet21K [14]. An experiment is conducted to measure the effect of this modification with different data augmentation techniques and the modified CNN backbone. The independent variable is the data augmentation techniques applied and the dependent variables are $top1Acc.$, $top5Acc.$ and categorical cross entropy loss. The $topKAcc.$ metric denotes the probability a queried sign is present in the top K results. equation (8.1) states the formal definition of the $topKAcc.$ metric [58]. The positives and negatives are counted if present in the top K results. The visual dictionary does not have negatives, only positives denoting the target sign. The $topKacc.$ thus measures the fraction of predictions where the target sign (True Positive) is in the top K predictions. This metric informally measures the dictionary performance, as it measures how often a user will find the input sign.

$$topKAccuracy = \frac{TruePositives + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative} \tag{8.1}$$

For each experiment the results of the epoch with the highest $top1Acc.$ is used. The data augmentations are stacked, meaning every experiment includes all previously added data augmentations. The data augmentation experiments are applied to half the training dataset with MFF configuration 4-MFFs-3f1c for two reason. First, the limited time and computing facilities did not allow for training on the full dataset. Secondly, the added value of data augmentation techniques should be more visible on smaller training sets, as overfitting is more present with smaller training sets.

Training is performed using the same parameters are used in the paper which introduces MFFs by Kopuklu et al. The configuration involves using the SGD optimiser with a batch size of 32 for

45 epochs with reduction of the initial learning rate of $1e3$ with a factor 10 after epoch 25 and 40. The result are shown below in table 8.1. The multilayer perception (MLP) was used in the paper as decision level fusion, however averaging was set as default in the source code [14][55].

The data augmentation methods are all adopted from the original paper[14], except GridMask [42]. GridMask is based on the deletion of regions of the input image [42]. GridMask was however removed due to the relatively low training $top1Acc$ of 92.64, compared to the validation $top1Acc.$ of 94.29%. This was a sign of underfitting, as the model was not able to adequately learn the training dataset. The hypothesis proved to be correct, as the $top1Acc.$ improved from 94.29% to 94.56%, whereas the training $top1Acc.$ increased to 95.96%.

Data augmentation	**Val Top1Acc. (%)**	**Top5Acc. (%)**	**Loss**
baseline with 50% of training data	82.94	97.70	1.1214
+ 20% horizontal shift and frame range 1	87.60	98.81	0.5851
+ 20% Padding and 20% Cropping	90.89	99.28	0.4185
+ 20 Degrees Rotate	90.94	99.29	0.3963
+ GridMask[42]	91.12	99.41	0.3196
+ 60% Training Data	92.13	99.51	0.2664
4-MFFs-3f1c baseline model [14]	*92.18*	*99.41*	*not available*
+ 75% Training Data	92.64	99.60	0.2523
+ 100% Training Data	92.86	99.65	0.2404
+ Multilayer Perceptron Classification	94.29	99.63	**0.1959**
+ Removed Gridmask	**94.56**	**99.69**	0.2031

Table 8.1: Data augmentation experiment results compared to the baseline model from the paper introducing MFFs [14]. The final model performs over 2 percentage point better than the baseline model in terms of $top1Acc.$

The results show an improvement of 2.38 percentage point over the results from the paper. Since the model is now trained on video's of gestures, conceptually close to the video's of NGT signs. In addition, the model is now trained with MFF's as input instead of RGB frames. The current weights $\theta Jester$ of h_{Jester} should therefore be closer to the optimal $\hat{h}_{NGT}$ weights than the $h_{ImageNet1K}$ weights. The next section will discuss the hyper parameter optimisation for the NGT training.

8.2 Similar and Different Signs Experiment

The lack of a test set makes it difficult to get confidence in the performance of the model. Machine learning has been highly successful in data-intensive applications but is often hampered when the data set is small [41]. This is especially true in the most extreme case, one-shot-learning, with only a single sample per class. The dataset provided contains a single sample per sign, making it a one-shot-learning problem. Without appropriate regularisation this could result in overfitting, which is a discrepancy between the performance on the training set and test set. A validation and test set are required to evaluate the performance of the model on untrained and unseen samples, to make sure the model is generalising and not just fitting the training data. Creating a validation and test set is however a time consuming process. To get a glimpse at the performance before the validation and test set were available an experiment using training data only was performed.

The model returns a ranked list of similar signs with respect to the input sign. A model fitted on the training set should yield a list with the input sign as head, followed by decreasingly similar signs. The performance of the model can thus be validated by analysing the location of similar and different signs. 25 triples consisting of two similar signs and one different signs are used to evaluate the performance of the model. The triples were facilitated by prof. dr. O.A. Crasborn. For each

triple, both similar signs have been used as input to determine the position of the other similar and different sign. The 25 triples result in 50 ranks for both a similar and different sign. In figure 8.1 a histogram showing the ranks gives an overview of the results. The blue bars denote the position of the similar signs and the orange bars denote the position of the different signs.

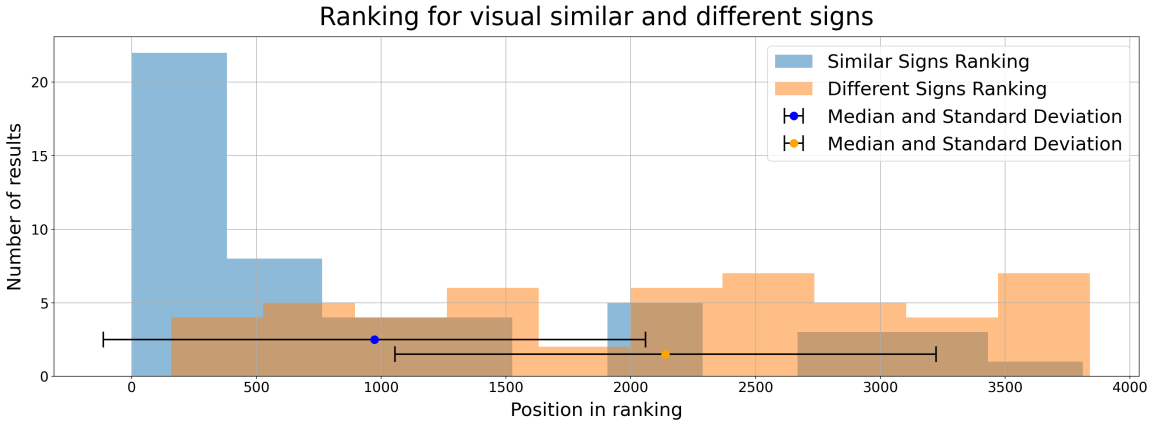

Figure 8.1: Ranking of visually similar and different signs out of 3846 signs. The blue bins denote the position of the similar sign to the input sign and the orange bins the position of the different sign.

A clear tail can be seen of similar signs which are labelled as similar, but ranked dissimilar by the model. Further manual analysis of why these signs are considered dissimilar is done by comparing the motion fused frames of the input sign, similar sign and second and third ranked sign. The difference between the similar signs seemed to lie in the key frame extraction, as the signs were not aligned in the motion fused frames.

In figure 8.2 this can be clearly seen, as the first frame of the input and similar sign are from different fractions of the sign. The optical flow frames are left out for readability reasons. In the first frame of the input sign the the hands are already located upwards, whereas in the first frame of the similar sign the hand are still in the neutral position. In the second frame the hand in the input frame are moving downwards, whereas the hands in the similar sign are moving upwards. This will result in opposite optical flow frames. Only the third and fourth frame are visually similar. The fourth frame is, however, the neutral position, which should be equal for each sign.

(a) Input sign (onzichtbaar) ranked 1st

(b) Similar sign (avond) ranked 3581th

(c) 2nd ranked sign (herfst) for given input sign

Figure 8.2: Input sign a with output sign b tanked 3581/3846 and secondly ranked sign c

The second ranked sign is visually similar to the input sign with aligned frames. In the first frame the hands are located upwards and in the second and third the hand are moving downwards and finally in the last frame the hands are in the neutral position. Especially the frames in the neutral position are problematic, as they yield no discriminative information.

This misalignment could be the reason behind the wrong ranking. Other mistakes were harder to understand, an example is the input and third ranked sign in figure 8.3. Here, the input sign has both a different position and movement compared to the third ranked sign, making it unclear why this sign is ranked second for the given input sign.

(a) Input sign (ga maar) ranked 1st

(b) 3rd ranked sign (been)

Figure 8.3: Input sign a and arguably visually dissimilar 3rd ranked sign b

8.3 Hyper parameter optimisation

After creating the validation and test the hyperparameters for training needed to be determined. These parameters can significantly affect the performance of the model. The goal of hyperparameter optimisation can be formalised as finding the optimal set of parameters λ^* that yields optimal model M^* which minimises the loss on the test set $L(X^{test}; M)$ [59]. Besides training parameters, these parameters can also include preprocessing parameters such as data cleaning or key frame extraction [59]. Finding the best hyperparameters is time consuming, because each combination of parameters requires training and evaluation of the model. An exhaustive search, where each combination of parameters is tested, is therefore often unfeasible.

Due to time constraints a sequential approach was used to find the best hyper parameters. A full grid search would result in $4 \cdot 3 \cdot 3 = 36$ runs, in comparison to $4+3+3 = 10$. Each iteration a different parameter was probed with different values. Each value was run three times due to variability in results caused by nondeterministic shuffling and processing of input data. Each run the parameter value with the best validation $top20Acc.$ was kept if it was an improved from the previous settings.

The parameters were initialised to the Jester transfer learning model produced in section 8.1.

Param	top1Acc.	top5Acc.	top10Acc.	Best top20Acc.
Number of Epochs [Frame Range=2, Dropout=0.50]				
Epochs 100 Run 1	5	13	18	30
Epochs 100 Run 2	4	19	31	44
Epochs 100 Run 3	3	9	18	32
Epochs 150 Run 1	6	19	28	45
Epochs 150 Run 2	14	22	35	52
Epochs 150 Run 3	7	19	33	51
Epochs 200 Run 1	6	23	36	51
Epochs 200 Run 2	12	26	38	51
Epochs 200 Run 3	8	29	39	54
Epochs 250 Run 1	7	15	32	49
Epochs 250 Run 2	7	29	35	52
Epochs 250 Run 3	6	24	34	52
Frame Range [Epochs=200, Dropout=0.50]				
Frame Range 1 Run 1	5	18	33	47
Frame Range 1 Run 2	6	23	32	48
Frame Range 1 Run 3	8	29	39	56
Frame Range 3 Run 1	10	28	38	59
Frame Range 3 Run 2	13	26	37	49
Frame Range 3 Run 3	11	31	45	60
Frame Range 4 Run 1	6	18	28	41
Frame Range 4 Run 2	6	14	28	45
Frame Range 4 Run 3	6	23	35	48
Dropout [Epochs=200, Frame Range=3]				
Dropout 0.40 Run 1	6	25	38	56
Dropout 0.40 Run 2	9	24	37	50
Dropout 0.40 Run 3	11	28	40	57
Dropout 0.60 Run 1	4	17	30	45
Dropout 0.60 Run 2	11	23	38	49
Dropout 0.60 Run 3	6	20	32	49
Dropout 0.70 Run 1	1	2	6	13
Dropout 0.70 Run 2	4	11	11	24
Dropout 0.70 Run 3	2	6	9	20

Table 8.2: Hyper parameter search results on a random subset of 300 signs of the NGT dataset with a validation set of 100 samples recorded in homely conditions by seven participants

The first parameter to optimise was the number of training epochs. Due to the one-shot learning approach these were set to values ranging from 100 to 250, with 50 epochs interval. Transfer learning from Jester to nvGesture, another gesture dataset with few samples per class, used 100 epochs with reducing the learning rate by 75% at 40% and 80% of the epochs. This learning rate reduction was extrapolated to the other number of epochs. The best performing value was found to be 200.

After reducing the learning rate at 80% of the epochs the validation *top*20*Acc.* did not improve anymore. The *top*20*Acc.* requires the top 20 most relevant signs. The model is trained as a classifier, where a probability distribution is given for the signs. The top 20 most relevant signs are selected by taking the 20 signs with the highest assigned chance. The number of epochs was therefore set to 160 with a learning rate reduction of 75% after epoch 80. Next, increasing the frame from two to three further improved the validation *top*20*Acc.*. Changing the dropout rate to 0.40, 0.50 or 0.60 did not result in an improvement. The final parameters configuration was 160 epochs, frame range 3 and 50% dropout.

8.4 Performance Evaluation on Test Set

Evaluating the model on a test gives confidence in performance on unseen and unoptimised samples. The samples in this test set have not been used for training nor evaluation. The test set should give confidence in the generalisability of the model, because the model can overfit on the training and validation data, but naturally not on unseen and unevaluated data. Where the evaluation set contains 100 samples, the test set contains 154 samples. The Zero Rule functions as a baseline, where the model would consistently output the majority class [60]. In a one-shot learning problem every class has one sample, meaning the Zero Rule baseline for *topK Acc.* with N classes equals $\frac{k}{N}$. This Zero Rule baseline functions as a naive baseline to check whether the model performs better than smart guessing. The weights of the model that performed best on the validation set during the hyperparameter optimisation process are used to evaluate the performance on the test set, the result are shown in table 8.3.

	top1Acc. (%)	**top5Acc. (%)**	**top10Acc. (%)**	**top20Acc. (%)**
Validation	11	31	45	60
Test	6	21	36	44
Zero Rule	0.33	1.67	3.33	6.67

Table 8.3: Performance of the best *top*20*Acc.* validation weights on the test set and the corresponding ZeroRule baseline. The model performs consistently better than the baseline.

The results show a consistent decrease in test performance compared to validation performance. This is a sign of validation set overfitting, where a model fits the validation data well, but performs worse on unvalidated test data. This could be partly caused by the extensive hyperparameter optimisation and making checkpoints of the best validation epoch. Due to the large amount of runs and epochs the model could fit the validation set well by chance. A larger validation set should make it less likely to fit the validation set by chance. However, due to the time consuming process of recording signs manually this is not feasible. Although the model performs worse on the test set than on the validation set, the model performs several factors better than the naive Zero Rule baseline. This gives confidence in the generalisability of the model on unseen and unvalidated data.

The recorded validation and test samples are stratified on the participants, thus each participant is present in both sets. One might argue the model is overfit on the participants and generalisability on unevaluated participants has not been defended. Further evaluation with a test set containing participants not present in the validation set should support this generalisability.

Lastly, the performance is analysed on a participant level. The results are shown in table 8.4. Participant 2 is a fluent NGT practitioner, but was sitting closely to the webcam making the hand movements not completely visible. Due to the low sample size it can not be concluded sitting close

in front of the webcam negatively impacts the recognition of signs. Moreover, some signs could be easier to retrieve than others by having unique, and therefore discriminative, properties which are relatively easy to recognise. The difference in performance could thus be caused by an uneven distribution of easy and hard to recognise signs over the participants. A more comprehensive test set with purposely different recording settings would give further insights in the correlation between recording setting and model performance. The philosophy behind this tool is however to let user visually search for signs in their natural setting. Adding far reaching constraints on their recording setting would undermine the use case of the tool.

	topKAccuracy (%)				
Participant	**# Samples**	**top1Acc.**	**top5Acc.**	**top10Acc.**	**top20Acc.**
Participant 1 (loc. 1)	33	6	12	33	42
Participant 1 (loc. 2)	30	3	40	47	57
Participant 2	23	0	17	22	30
Participant 3	18	11	11	28	39
Participant 4	18	0	17	28	33
Participant 5	16	6	13	31	38
Participant 6	16	19	31	62	69

Table 8.4: Performance on test set per participant using the weights of the best validation run

This section evaluated the performance of the model using a test set with video's of signs recorded in a homely setting. The performance is quantified using the *top*1*Acc.* metric, which measures the probability of the queried sign to be present in the top K results. This does not take into account the relevance of the other $K-1$ retrieved signs. The next section will quantify the relevance of all retrieved signs to better quantify the performance from an IR view.

8.5 Sign Language Similarity NDCG

So far, the performance has been measured using the *top*1*Acc.* metric, which only takes into account the occurrence of a single target label in the top K results. An IR system should provide the documents which are most relevant to the information need of the user, ranked on relevance. user translate their information need to a video recording of them performing a sign, which is then matched with all other recordings of signs in the database. To correctly measure the relevance of all retrieved documents, as well as taking the order into account, the normalised discounted cumulative gain metric (NDCG) is used [13].

The same model as in section 8.4 is used, which is the best performing model of the hyperparameter search. Both the validation and test set are evaluated using the $NDCG@K$ metric for the same K used for the *topKAcc.* evaluation, meaning 1, 5, 10 and 20. In addition, a naive baseline score is computed using 1000 random rankings on the complete 300 input images, resulting in 300,000 random $NDCG@K$ scores. The results are shown in table 8.5.

Dataset/Performance (%)	**NDCG@1**	**NDCG@5**	**NDCG@10**	**NDCG@20**
Validation	68	61	59	59
Test	63	57	57	58
Baseline	40.3	43.8	45.6	48.0

Table 8.5: Normalised Discounted Cumulative Gain on NGT validation and test dataset with baseline.

table 8.5 shows a consistent improvement over the baseline. Additionally, the performance of on the validation set is slightly better than the performance on the test set. This is again a sign of validation set overfitting. For higher K values the difference between the baseline and results becomes smaller, indicating a lower relevance when more results are shown. In absolute terms, the difference between the baseline and $NDCG$ score becomes smaller than K increases. This indicates a reduction in performance when more results are shown. The constructed relevance score applied using the NDCG shows the model is learning to retrieve relevant results above the baseline. This suggests the model does not only function as a classification model, but also as an IR model.

To get a better understanding of what the model is picking up the NDCG score per phoneme is analysed. Besides the validation and test set a baseline is computed using 1000 random ranking of the complete 300 training samples. The result are shown below in section 8.5.

	NDCG@20 (%)				
	Location	**Movement**	**Strong Hand**	**Weak Hand**	**Handedness**
Validation	63	39	29	28	50
Test	64	37	26	29	50
Baseline	58.5	35.4	22.7	24.1	45.3

Table 8.6: Phoneme $NDCG$@20 for each on both the validation and test set and the corresponding baseline scores

Both the validation and test $NDCG$@20 score consistently better on all phonemes. There is no clear dominant phoneme the model picks up best. Given MFF's use optical flow frames one might expect movement to be the dominant phoneme, however this is not the case. Strong and weak hand shapes are surprisingly picked up, which was not expected given the low video quality.

8.6 MFFs and RGB frames as Document Represenation

The superiority of MFF's over RGB frames has been shown in a sign classsification setting by Kopuklu et al. on the Jester dataset [14][54]. To measure the added value of appended optical flow frames to RGB frames the performance of different MFF configuration is measured. With a *N-MFF-XfYc* configuration only the number of optical flow frames X is modified. The number of RGB frames Y has not been modified in the original paper. Due to time constraint the number of segments N is not modified either. The result are shown below in table 8.7.

Performance/MFF Config.	**4-MFF-0f1c**	**4-MFF-1f1c**	**4-MFF-2f1c**	**4-MFF-3f1c**
Number of Frames	12	20	28	36
Jester Val@1	93.22	94.42	94.42	**95.52**
NGT Val top20Acc.	46	59	57	**60**
NGT Val NDCG@20	51	58	57	**59**
NGT Test top20Acc.	36	**50**	42	44
NGT Test top10Acc.	25	35	33	**36**
NGT Test NDCG@20	55	57	**58**	**58**
NGT Test NDCG@10	55	56	**58**	57

Table 8.7: Performance of different motion fused frame configuration on Jester and NGT dataset showing the added value of appended optical flow frames to RGB frames

The results are surprising for several reasons. Firstly, the added value of the first optical flow frame is high with a validation $top1Acc.$ jump from 93.22% to 94.42% on the Jester dataset. Any

further added optical flow frames do not seem to heavily impact the performance. Future research should clarify whether the improvement from two to three optical flow frames is a statistical outlier. On the NGT subset of 300 all MFF configurations perform consistently better than RGB frames on every metric. There is however no consistently better performing MFF configuration on the test set.

Arguably, more optical flow frames could result in more overfitting in a one-shot learning problem given the higher number of input frames. However, this does not seem to be the case. The added value of the first optical flow frame gives both an improvement in $top1Acc$ and $NDCG$, indicating superiority of MFF over RGB frames. Additional optical flow frames do however not consistently improve the top1Acc and NDCG, leaving the marginal added value of additional optical flow frames unknown.

8.7 RQ3 Conclusion

This chapter discussed the performance of the model in both a classification and IR setting. Performance quantification using the $top1Acc.$ and $NDCG@K$ metric showed a consistent improvement over the baseline. MFF's proved to be superior over RGB frames in both a classification and IR setting, however the added value of additional optical flow frames is unknown.

Chapter 9

Sign Language Key Frame Extraction

This chapter explains the theoretical framework and implementation of a domain specific sign language key frame extraction algorithm. The first sections described the theory and implementation of the proposed method. Next, a performance comparison is done between uniform sampling and the proposed key frame extraction method.

9.1 Sign Language Structure

Manual signs phonemes consist of location, movement and hand/arm configuration [15]. These phonemes form a sign and are thus the discriminative properties of a sign. When extracting the key frames of a sign the frames which best capture those properties should be selected. A method based on the body keypoints detection tool OpenPose[12] could capture solely hand movement and thereby forming a baseline for a domain specific key frame selection method for sign language. Together with the phonetic structure of a sign the following key frame extraction method is proposed.

The first step consists of identifying the dominant hand. This is done by selecting the hand that has the largest movement. The movement of each hand quantified by summing the absolute values for both the x an y deltas. Next, the start and end of a sign need to be unidentified. Every sign starts and ends in the neutral position. Firstly, the hand moves to the start location, followed by an optional movement and ends with a movement to the neutral position. As each sign has this start and end movement from/to the neutral position these movements are not a discriminative property of a sign and should therefore be left out. The frame the hand reaches the start location can be detected by a vertical movement followed by stagnation of movement at the start of the video. OpenPose can detect this by a local optimum at the start of the video for the y delta. For the movement from the end location to the start location the same applies, but at the end of the video. Movements between the start and end location can thus be detected by local optima between the start and end location. Due to noise local optima are only taken into account when their absolute value is bigger than 1% of the resolution and both the two deltas on the left and right are both consistently decreasing or increasing. If no local optima are found between the start and end location, which is the case for signs without a movement, uniform sampling is applied between the start and end location. If there are more than two optima the two biggest optima are selected, thereby capturing the largest movements. When a single optima is found uniform sampling is applied between the optima and the start or and frame, depending on which distance is the largest. An example of the key frame selection process is visualised in figure figure 9.1 for the sign pensioen-B (Pension-B).

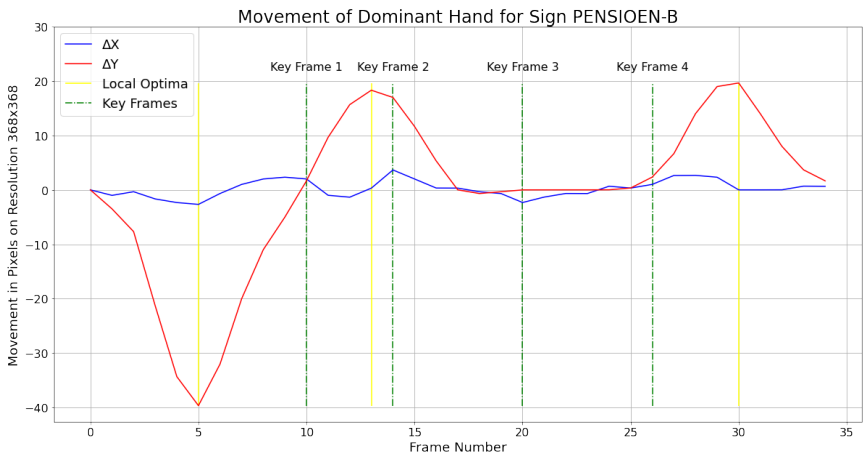

Figure 9.1: Key Frame Extraction Process using local optima of hand movement deltas in pixel extracted with OpenPose

The yellow vertical lines denote local optima and green denote the selected key frames. The movement from and to the neutral position at the start and end of the sign can be clearly seen. The first frame for which the movement in the y direction is below the threshold of 1% of the resolution is selected as start or end frame. The start frame is selected after the local optima and the end frame before the local optima, because at the start the hand moves from the neutral position to the start position and at the end the hand moves from the end position to the neutral position. Because the hand movement is near zero in the start and end key frames the hand shape and location should be captured. In frames with high hand movement motion blur can occur, making the hand shape hard to detect. Key frame 2 is selected because it is a local optima between the start and end frame. The frame after the local optima is chosen to capture the maximum movement in the optical flow frames. Since there are no more local optima left key frame 3 is selected by uniform sampling between key frame 2 and 3, because the distance in frames is larger between key frame 2 and 4 than between 1 and 2.

This combination of static start and end frames in combination with capturing movement between these two frames should both capture spatial information: hand shape and location, as well as temporal information: movement. To compare the performance of this key frame extraction method to the uniform sampling key frame extraction method both methods are used to construct a training, validation and test dataset. The best performing hyper parameters found in section 8.3 are used for training the model on OpenPose key frame extracted MFF's. As with uniform sampling, the weights of the best *top*20*Acc.* epoch out of three runs is selected. The validation and test performance for uniform sampling and OpenPose key frame extraction are compared in table 9.1.

	Precision				**NDCG**			
	@1	**@5**	**@10**	**@20**	**@1**	**@5**	**@10**	**@20**
Uniform Sampling Val	11	31	45	60	68	61	59	59
OpenPose KFE Val	14	31	37	54	69	59	57	58
Uniform Sampling Test	6	21	36	44	63	57	57	58
OpenPose KFE Test	16	31	46	55	69	60	59	59

Table 9.1: Comparison between uniform sampling and custom OpenPose key frame extraction method performance on the validation and test set.

Surprisingly, the OpenPose key frame extraction method performs better on the test set than on the validation set. This might be partly explained by the absence of hyperparameter optimisation,

eliminating a source of validation set overfitting. The best epoch out of 160 epochs over three runs is selected, which would still allow for validation set overfitting. The results do give confidence in the performance of the domain specific key frame extraction method. Especially the *top*1*Acc.* of 16% is impressive, given a Zero Rule baseline of just 0.33%.

9.1.1 Improvements

This key frame extraction method does rely on the detection of the wrists by Openpose and could be further improved in a variety of ways. Firstly, the low quality of videos recorded with webcams caused OpenPose to occasionally not detect the wrists correctly or not at all. Missing wrist coordinates were interpolated. The histogram in figure 9.2 shows the distribution of missing wrist coordinates. In some cases over half the wrist coordinates were missing. This percentage is also a strict minimum, as only missing values between the first and last identified wrist coordinates are counted. There could be missing wrist coordinates before the first and after the last wrist coordinate. Whereas OpenPose predicts the position of the wrist within a margin of half the head size with a precision of 84.7%[61] on the MPII test set current state of the art models achieve a precision of 91.2% [62]. An improved wrist detector would reduce the amount of incorrect or missing wrist detections, resulting in more accurate movement deltas and therefore better key frame detection. Next, several hyperparameters have been heuristically set, such as start and end key frame thresholds for static frames, local optima thresholds and frame offset for intermediate local optima. These parameters should empirically be determined for optimal performance. Lastly, OpenPose could provide additional features as input to the model. For example the maximum height of the hands and handedness could be determined using OpenPose, as well as the start and end coordinates. OpenPose also allows for detailed hand keypoint estimation, where coordinates of each phalanx of each finger are provided. This does however require high quality videos in a high resolution to allow for sufficient detail when cropping the hand out of a frame. Features based on hand keypoints could for example be used for hand shape estimation. The potential of body and hand keypoint estimation is arguable large for sign language recognition. Improved wrist estimation, empirical hyperparameter optimisation together with additional body and hand keypoint features could push the performance of sign language recognition to the next level.

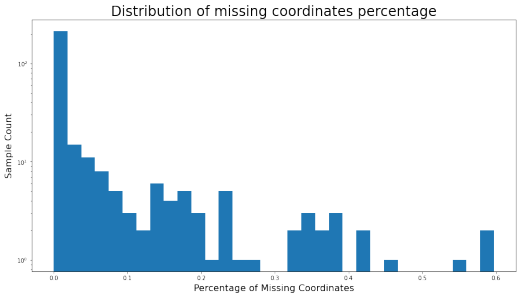

Figure 9.2: Distribution of percentage missing coordinates per record, showing OpenPose often had difficulty identifying all wrist coordinates

9.2 Learning to Rank

The OpenPose key frame extraction method did improve both the *precision* and $NDCG$ score, thereby improving both the capabilities of the visual dictionary for finding signs and retrieve visually similar signs. The $NDCG$ score is however just marginally improved, especially the $NDCG$@20 increment from 58 to 59 is minimal. This could be explained by the learning objective. The model is not trained to retrieve similar signs, but to classify an input sign. A L2R (Learning to Rank) approach would harmonise the learning objective with the IR use case. L2R is about performing ranking using machine learning techniques [63]. To increase the $NDCG$ score a pointwise approach is applied to improve the relevance of retrieved results. This approach creates a typical supervised learning task where an input variable X is mapped to an output value y [63]. Specifically, the learning objective becomes the task of predicting the relevance score y for each sign in the dataset for a given input sign X. In comparison to the classification approach the target is a vector with a target value in the range $[0, 1]$ for every sign in the dataset, instead of a single target label. To test the effectiveness of this modified learning approach the model is trained using the pointwise

approach. During development the model did not fit the training data well when using the same hyper parameters as for classification training. The number of epochs is therefore increased from 200 to 500, the second dropout reduced from 0.50% to 0.25% and the learning rate is not reduced from the initial value of 0.001. This reduced regularisation caused the training metrics to improve and plateau, indicating the model was actually fitting the training data. This resulted in a $top20Acc.$ of 5% and $NDCG@20$ of 49%, which is approximately naive baseline performance. This could be partially explained by the binary cross entropy loss function which does not take into account the relevance of a target when computing the loss. More specifically, prediction errors for low relevance signs are taken equally into account as errors in high relevance signs. In an IR setting only the top ranked results are actually retrieved by the user. A prediction error for a highly relevant sample could result in the user not retrieving this relevant sample. A prediction error for a low relevance sample would rank the sample less low and would only be problematic if the sample would actually be retrieved by the user. Consider a sign ranked 250th, if the user only retrieves the top 20 ranked signs it would not matter if the sign is ranked precisely 250th or 200th, or even 100th. A relevant sign ranked 10th would actually need to be in the top 20 ranked signs to be retrieved by the user, allowing for a smaller error margin.

To make the loss function rank aware a DSL (dynamic loss scaling) method is introduced which scales the loss of prediction with respect to the relevance score. The formula used is $\frac{1}{max(epsilon,relevance)}$ where the epsilon prevents the loss scale from zero division and becoming to large. DSL scales the loss of signs with respect to the relevance, where a higher relevance results in a higher loss scaling. This makes the loss function rank aware by assigning weights to the loss of each sign based on the relevance, where relevant signs are weighted higher than less relevant signs. The loss scale with different epsilon values is shown in figure 9.3. The best performing epsilon value was, surprisingly, 0.01 which scaled the loss of the target label by a factor $\frac{1}{max(0.01,1-1}) = 100$. This DSL method resulted in a non trivial $top20Acc.$ improvement from 5% to 36% and an $NDCG@20$ improvement from 49% to 72%. This DSL method improves both improves the classification and IR performance.

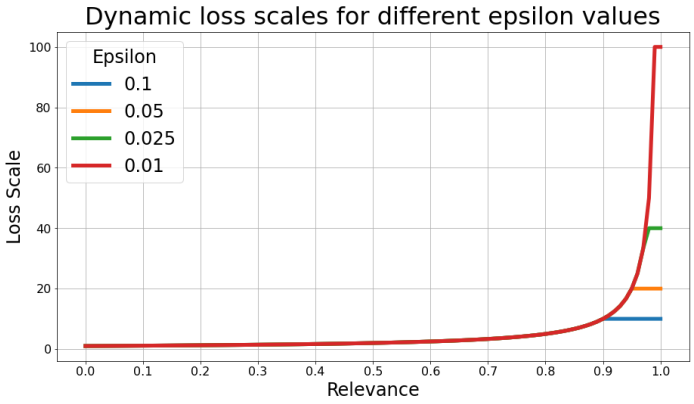

Figure 9.3: Dynamic loss scaling with respect to relevance computed by $\frac{1}{max(epsilon,relevance)}$ for multiple epsilon values.

A second improvement was introduced to improve the classification performance of the L2R approach. This method multiplies the loss of the target sign after DSL is applied and is called Target Loss Scaling (TLS), because it scales the loss of the target sign. Since several signs could have identical labels multiple sign could have identical relevance targets of 1.0 with identical loss scaling. TLS scales the loss scale of the target sign to distinguish it from other signs with identical labels. This improves the importance of correctly predicting the relevance of the target sign in the loss function. A loss scaling of two improved the $top20Acc.$ with 8 percentage point from 36% to 44%, resulting in the same $top20Acc.$ as the uniform sampling baseline. The $NDCG@20$ was reduced by 1 percentage point from 72% to 71%. This TLS method shows how the $topKaccuracy$ can be improved in a L2R training method by scaling the loss of the target sign with a minimal $NDCG@K$ reduction.

This pointwise L2R approach with DSL and TLS shows how a pointwise learning to rank approach improves the relevance of results for sign language learners whilst also having adequate classification performance. The DSL espilon and TLS scale provide flexibility between $topKaccuracy$ and $NDCG@20$ performance. This allows to balance the capabilities of the visual dictionary for finding

single signs and retrieving overall relevant signs.

	topKaccuracy (%)				**NDCG (%)**			
	@1	**@5**	**@10**	**@20**	**@1**	**@5**	**@10**	**@20**
Uniform Sampling Test	6	21	36	44	63	57	57	58
OP KFE Test	16	31	46	55	69	60	59	59
OP KFE L2R Test	1	1	3	5	38	45	47	49
OP KFE L2R DLS(0.01) Test	4	16	22	36	66	68	69	72
OP KFE L2R DLS(0.01) TLS(2) Test	6	19	31	44	67	68	69	71

Table 9.2: Performance comparison between uniform sampling, OpenPose Key Frame Extraction (OP KFE) and a learning to rank (L2R) approach with additional Dynamic Loss Scaling (DLS) and Target Loss Scaling (TLS).

9.3 RQ4 Conclusion

The proposed key frame extraction method is specifically designed to capture discriminative frames in a recording of performed sign. OpenPose is applied to detect hand movement. A method based on the structure of sign is introduced which detects the start location and handshape, optional movements and the end location and handshape. This method performs better than uniform sampling, giving confidence in the effectiveness of this key frame extraction method.

Additionally, a pointwise learning to rank method (L2R) is introduced whose learning objective is to predict the relevance of each sign in the dataset for a given input sign. Dynamic loss scaling (DSL) and target loss scaling (TLS) are introduced to improve the classification performance. L2R offers the flexibility to balance the capabilities of the visual dictionary between finding single signs and overall relevance of the retrieved signs.

Discussion

The Centre for Language and Speech Technology at Radboud University requested the development of a visually searchable dictionary for sign language learners. This thesis aimed to fulfil this request by exploring the linguistic and technological challenges of a visually searchable NGT IR system. Such a system eliminates the need for the manual labelling of sign properties and provides user with the ability to search a dictionary in their native language. This research has been subdivided in four research questions of which the limitations and potential areas for future research are discussed below.

RQ1: Requirements Engineering

Firstly, the interviewees indicated the need for the functionality to retrieve semantically close words, which could be implemented with off-the-shelf word vectorisation models. Other requirements concerned the use case, results layout, additional information on signs and video processing. These requirements are not, or only partially, implemented in the available online NGT dictionary. This illustrates the importance of user involvement in software development. This results in a mismatch between user requirements and realised system which could be prevented with proper RE.

Conventional requirements engineering development requires several iterations in which stakeholders can give feedback on visual progress to further modify, add or remove requirements. Future work could implement the derived user stories and perform a second iteration to further improve the requirements for a visually searchable dictionary with a minimal prototype. The current lack of a minimal prototype can impact the requirement elicitation process by keeping the visual dictionary abstract. The identified requirements could be modified, removed or added in future RE iterations, meaning the identified requirements are not final.

RQ2: Sign Similarity

Secondly, a similarity quantification for sign language is introduced based on interviews with stakeholders in the deaf community. These interviews provided a clear picture on how sign language learners perceived similarity in sign language. Locations can be clustered in visually similar areas and movement is characterised by the axis it follows. Interviewees did however have difficulty with defining similarity between handshapes. Therefore, the handshape properties introduced by van der Kooij [3] were used to define generic handshape properties. Further research could focus on the discriminative value of each individual handshape property. The difficulty for defining handshape similarity could indicate handshape is the most complex phoneme. This would be in line with prior research on people's ability to remember signs, which showed that most phonological errors were made in the handshape [21].

Nonmanual phonemes were not included in the sign similarity metric, because of time constraints and detection difficulty. Including nonmanual phonemes could further improve the similarity quantification. The five similarity scores are uniformly weighted, but it is likely this does not accurately reflect the sign language learners view on sign similarity. Movement is by many interviewees viewed as a visually dominant, followed by location, whereas handshape is more subtle and therefore visually submissive. Handedness would arguably be assigned the highest weight, since confusing one-handed signs with two-handed signs was considered rare. Mistakes between symmetrical and a-symmetrical two-handed signs is unknown and should be further investigated. Future research could thus consists of empirically determining the weights of each phoneme in the sign similarity formula.

Both RQ1 and RQ2 relied on the same interview group of seven people. The small sample size limits the generalisability and reproducibility of the results. Furthermore, for six out of the seven interviewees NGT was their primary sign language, making the generalisability of the results to other

sign languages unknown. Lastly, the four that were enrolled in a formal sign language education program all studied at the same university, which could have resulted in a biases due to the study program.

RQ3: Motion Fused Frames in Information Retrieval

Thirdly, MFF's are shown to be superior to conventional RGB frames in an IR setting. This is in line with prior research where MFFs performed better than RGB frames in a sign classification problem. Marginal value of additional optical flow frames is unsure on the other hand, both in a classification setting and an IR setting. Intuitively, the additional optical flow frames should provide more information and thus result in better performance. Prior research showed a concistent improvement with additional optical flow frames in a classification setting, which is in contrast with the results in RQ3. Conversely, a combination of information saturation and overfitting could however be the cause of the unclear impact of additional optical flow frames. Research on the impact of additional optical flow frames would give a better understanding on MFF's. Due to time constraints the effect of the number of segments and RGB frames has not been not determined. Future research could provide further insights into MFF's by determining the effect of these configuration options.

When manually reviewing the results of the visual dictionary visually dissimilar signs are retrieved. The visual dictionary is based on a neural network, which makes explainability challenging. This lack of explainability plays a central role in machine learning and there is often a trade-off between accuracy and explainability. Neural networks often offer the highest accuracy according to a recent study, while also having the lowest explainability [64].

RQ4: OpenPose Key Frame Extraction

Lastly, a domain specific key frame extraction method for sign language based on OpenPose wrist detection proved dominant over uniform sampling. This method can detect the start and end position of a sign as well as movement between these positions. In many cases OpenPose could not correctly detect all wrist positions between the start and end position, indicating the first and last detection of the wrist could be incorrect too. The implementation of the solution is therefore limited by imperfections in the wrist detection. Improving the wrist detection could potentially further improve the performance. Additionally, features based on body and hand keypoints could further improve the performance of the model. Human evaluation of the selected key frames could give insights in improvement strategies and quality of the method. The development of a domain specific key frame extraction algorithm for sign language could function as a fundamental tool in the field of sign language recognition.

Lastly, both implementations in RQ3 and RQ4 used a validation and test set recorded in home environments that used built-in laptop webcams. The quality of these recordings is low compared to the studio recordings in the dataset. The decision for built-in laptop webcams is made to resemble real world usage as close as possible. Improved webcam quality reduces noise in recordings which would improve the optical flow frames and wrist detection. Future results could therefore differ due to improved camera quality.

This research made the first explorative steps in the linguistic and technological challenges of introducing a visual dictionary that supports sign language learning. The current results have their limitations, but help provide the foundation on research in the field of visual dictionaries for sign language learners with a variety of options on future work.

Conclusion

This thesis explored the linguistic and technological challenges of a visually searchable sign language dictionary that supports sign language learning. This can allow a students to record a video of themselves performing a sign or generic gesture and retrieve the most similar signs via the tool. By doing so, it gives sign language learners the option to use NGT to search for NGT signs, instead of searching on Dutch translations of NGT signs or sign parameters.

RQ1: Requirements Engineering

To develop this solution, first, interviews were held with sign language learners, a teacher, an interpreter and a PhD candidate researching sign language to determine which requirements such a dictionary should have. Requirements identified consisted of functional requirements regarding the retrieved signs and additional information and visual requirements regarding the design. These requirements were formulated as user stories and should clarify the wishes of sign language learners from a visual dictionary. The learning potential of a visual dictionary can be improved by implementing the functionality and adopting the design defined in the user stories.

RQ2: Sign Similarity

After obtaining the requirements for the visually searchable dictionary an evaluation metric was constructed to quantify the relevance for sign language learners of retrieved signs. The relevance of the retrieved signs was defined by their similarity to the input sign. This required the quantification of perceived sign similarity by sign language students. Sign similarity is defined using the phonemes location, hand shape, handedness and hand shape. A taxonomy tree is constructed for location, hand shape and handedness where similarity is defined using the node distance. Defining generic hand shapes was considered difficult. For this reason the generic hand shape properties by van der Kooij were used as labels for hand shapes [3]. Similarity between hand shape is defined by the fraction of overlapping hand shape properties. Combining all of this results in five similarity scores for the location, movement, strong hand shape, weak hand shape and handedness, whose mean is the final similarity score. This relevance score is applied to the $NDCG$ metric for evaluating the relevance of retrieved signs by the visually searchable dictionary. Sign language learners will benefit from this similarity quantification by providing developers with a means to optimise the relevance of retrieved results of a visual dictionary.

RQ3: Motion Fused Frames in Information Retrieval

With the evaluation metric defined the actual visual dictionary could be developed. The work of Kopuklu et al. was used as a basis [14]. This model uses MFFs, where optical flow images are appended to RGB frames to both capture spatial and temporal features. To correctly evaluate the performance a validation and test is created of respectively 100 and 153 different signs that were recorded in home environments with conventional webcams to closely imitate usage in practice. Compared to RGB frames, MFFs improved the $top1Acc.$ from 36% to 44% and the $NDCG@20$ from 55% to 58% on the test set. This improvement in performance indicate the appended optical flow frames containing temporal information yield added value and improve the relevance of results for sign language learners. When varying the number of optical flow frames no consistent improvement is perceived, leaving the added value of additional optical flow frames undetermined. The performance dominance of MFFs over RGB frames in a sign IR setting provides confidence in the appropriateness of MFFs as data representation for sign recordings. This could make software developers consider

to adopt this novel data representation leading to inimproved performance of supportive tools for sign language learners.

RQ4: OpenPose Key Frame Extraction

Lastly, a domain specific sign key frame extraction method is introduced. This method relies on the generic structure of a sign. A sign is characterised by the starting location, optional movement, end location and hand shape [15]. With OpenPose the left and right wrists are detected in each frame and used to determine the hand movement between frames. The start and end location, as well as optional movement, are identified using local optima of movements. This domain specific sign key frame extraction method should capture both spatial and temporal information and thus provide information on the location, movement and hand shape. The method resulted in a $top20Acc.$ improvement from 44% to 55% and an $NDCG@20$ improvement from 58% to 59%. This improvment in performance illustrates the added value of the introduced key frame extraction method over uniform sampling. This key frame extraction method could provide the theoretical foundations of salient frame detection and definition in sign language recordings. Improving the key frame extraction will increase the discriminative value of selected frames and thus the discriminative value of the document representation. This will simplify sign language classification and information retrieval tasks. Sign language learners will benefit by improved relevance of retrieved signs.

This research explored the challenges of introducing a visual NGT dictionary for sign language learners. Fundamental insights have been obtained in both the linguistic and technological aspects of this engaging topic in order to get a step closer to the introduction of a publicly available visual dictionary for sign language. These findings should help lay the foundation for research on visual dictionaries and provide academics with ideas for future research. It is fascinating to discover how the upsurge of deep learning techniques can help provide novel tools that enhance the development and societal inclusion of people with auditory disabilities, whilst at the same time providing valuable insights from an academic point of view.

Bibliography

[1] M. Wijkhuizen, "Thesis github repository." https://github.com/MarkWijkhuizen/Supporting-Sign-Language-Learning-With-a-Visual-Dictionary.

[2] I. Zwitserlood, "Sign language lexicography in the early 21st century and a recently published dictionary of sign language of the netherlands," *International Journal of Lexicography*, vol. 23, no. 4, pp. 443–476, 2010.

[3] E. Van der Kooij, "Phonological categories in sign language of the netherlands," *The Role of Phonetic Implementation and Iconicity. LOT, Utrecht*, 2002.

[4] M. Fragkiadakis, V. Nyst, and P. van der Putten, "Signing as input for a dictionary query: Matching signs based on joint positions of the dominant hand," in *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pp. 69–74, 2020.

[5] O. Alonzo, A. Glasser, and M. Huenerfauth, "Effect of automatic sign recognition performance on the usability of video-based search interfaces for sign language dictionaries," in *ASSETS '19: The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 56–67, 2019.

[6] Y. Min, Y. Zhang, X. Chen, and X. Chai, "An efficient pointlstm for point clouds based gesture recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5760–5769, 2020.

[7] A. Elboushaki, R. Hannane, K. Afdel, and L. Koutti, "Multid-cnn: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in rgb-d image sequences," *Expert Systems with Applications*, vol. 139, 2020.

[8] Z. Lu, S. Qin, X. Li, L. Li, and D. Zhang, "One-shot learning hand gesture recognition based on modified 3d convolutional neural networks," *Machine Vision and Applications*, vol. 30, no. 7-8, pp. 1157–1180, 2019.

[9] M. Fragkiadakis and P. van der Putten, "Sign and search: Sign search functionality for sign language lexica," in *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, (Virtual), pp. 23–32, Association for Machine Translation in the Americas, Aug. 2021.

[10] H. Zhao, W.-J. Wang, T. Wang, Z.-B. Chang, and X.-Y. Zeng, "Key-frame extraction based on hsv histogram and adaptive clustering," *Mathematical Problems in Engineering*, vol. 2019, pp. 1–10, 09 2019.

[11] W. Wolf, "Key frame selection by motion analysis," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1228–1231, 1996.

[12] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2020.

[13] F. Radlinski and N. Craswell, "Comparing the sensitivity of information retrieval metrics," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 667–674, 2010.

[14] O. Kopuklu, N. Kose, and G. Rigoll, "Motion fused frames: Data level fusion strategy for hand gesture recognition," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2018, pp. 2184–2192, 2018.

[15] W. C. Stokoe, "Sign language structure," *Annual Review of Anthropology*, vol. 9, no. 1, pp. 365–390, 1980.

[16] NOS, "Steun in kamer voor officiële erkenning gebarentaal: 'we zijn er nog niet na irma'." https://nos.nl/artikel/2346338-steun-in-kamer-voor-officiele-erkenning-gebarentaal-we-zijn-er-nog-niet-na-irma, Sept. 2020. [Online; accessed 27-June-2020].

[17] R. Cokart, T. Schermer, C. Tijsseling, and E. Westerhoff, "In pursuit of legal recognition of the sign language of the netherlands," in *The Legal Recognition of Sign Languages: Advocacy and Outcomes Around the World*, pp. 161–175, Multilingual Matters, 2019.

[18] Nederlands Gebarencentrum, "Online Gebarenwoordenboek," June 2021.

[19] T. Schermer and C. Koolhof, *Van Dale basiswoordenboek Nederlandse gebarentaal*. Bunnik: Nederlands Gebarencentrum, eerste editie, zesde oplage. ed., 2018.

[20] I. P. Association and I. P. A. Staff, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[21] U. Bellugi, E. S. Klima, and P. Siple, "Remembering in signs," *Cognition*, vol. 3, no. 2, pp. 93–125, 1974.

[22] R. Rastgoo, K. Kiani, and S. Escalera, "Sign language recognition: A deep survey," *Expert systems with applications*, 2021.

[23] M. Aktas, B. Gokberk, and L. Akarun, "Recognizing non-manual signs in turkish sign language," in *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA) Istanbul, Turkey 2019 Nov. 6 2019 Nov. 9*, pp. 1–6, IEEE, 2019.

[24] N. Ejaz, I. Mehmood, and S. Wook Baik, "Efficient visual attention based framework for extracting key frames from videos," *Signal Processing: Image Communication*, vol. 28, no. 1, pp. 34–44, 2013.

[25] S. Zhang, Z. Zhu, and Z. Hu, "Sign language recognition based on key frame," *IOP Conference Series: Earth and Environmental Science*, vol. 252, no. 3, 2019.

[26] S. Jadon and M. Jasim, "Unsupervised video summarization framework using keyframe extraction and video skimming," in *5th International Conference on Computing Communication, 2020 I. E. E. E. and Automation (ICCCA) Greater Noida, India 2020 Oct. 30 2020 Oct. 31*, pp. 140–145, 2020.

[27] G.-H. Song, Q.-G. Ji, Z.-M. Lu, Z.-D. Fang, and Z.-H. Xie, "A novel video abstraction method based on fast clustering of the regions of interest in key frames," *AEUE - International Journal of Electronics and Communications*, vol. 68, no. 8, pp. 783–794, 2014.

[28] Radboud University, "Centre for language and speech technology." https://www.ru.nl/clst/, Oct. 2021.

[29] Q. K. Shams-Ul-Arif and S. A. K. Gahyyur, "Requirements engineering processes, tools/technologies and methodologies," *International Journal of reviews in computing*, vol. 2, no. 6, pp. 41–56, 2009.

[30] K. Harbison and K. McGraw, "User-centered requirements: The scenario-based engineering process," 1997.

[31] S. Malviya, M. Vierhauser, J. Cleland-Huang, and S. Ghaisas, "What questions do requirements engineers ask?," in *2017 IEEE 25th International Requirements Engineering Conference (RE)*, pp. 100–109, 2017.

[32] IEEE, "Ieee recommended practice for software requirements specifications," 1998.

[33] K. E. . Wiegers, "Software requirements : practical techniques for gathering and managing requirements throughout the product development cycle," 2003.

[34] G. Lucassen, F. Dalpiaz, J. M. E. M. Van Der Werf, and S. Brinkkemper, "Forging high-quality user stories: towards a discipline for agile requirements," in *IEEE International Conference on Requirements Engineering*, pp. 126–135, 2015.

[35] M. Cohn, *User stories applied: For agile software development.* Addison-Wesley Professional, 2004.

[36] R. B. Grady, "An economic release decision model: Insights into software project management," in *Proceedings of the Applications of Software Measurement Conference*, pp. 227–239, 1999.

[37] A. M. Davis, *Software requirements: objects, functions, and states.* Prentice-Hall, Inc., 1993.

[38] U. Bellugi and S. Fischer, "A comparison of sign language and spoken language," *Cognition*, vol. 1, no. 2-3, pp. 173–200, 1972.

[39] H. Wang, A. Stefan, S. Moradi, V. Athitsos, C. Neidle, and F. Kamangar, "A system for large vocabulary sign search," in *Trends and Topics in Computer Vision*, pp. 342–353, 2010.

[40] F. W. Lancaster, A. J. Warner, and B. Frohmann, "Information retrieval today.," *Journal of documentation*, vol. 51, no. 1, pp. 76–77, 1995.

[41] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020.

[42] P. Chen, S. Liu, H. Zhao, and J. Jia, "Gridmask data augmentation," *arXiv*, 2020.

[43] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Conference Proceedings*, pp. 6023–6032, 2019.

[44] M. Nixon and A. Aguado, *Feature extraction and image processing for computer vision.* Academic press, 2019.

[45] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Lecture Notes in Computer Science*, pp. 25–36, Springer, 2004.

[46] Nederlands Gebarencentrum, "De gesproken component in ngt." https://www.gebarencentrum.nl/gc, Oct. 2021.

[47] Nederlands Gebarencentrum, "De nederlandse gebarentaal." https://www.gebarencentrum.nl/Wat

[48] E. Keen, *A primer in phenomenological psychology.* University Press of America, 1975.

[49] R. S. Valle, M. King, and S. Halling, "Existential-phenomenological perspectives in psychology: Exploring the breadth of human experience," in *An introduction to existential-phenomenological thought in psychology*, pp. 3–16, Springer, 1989.

[50] B. J. M. Emans, *Interviewen: theorie, techniek en training.* Stenfert Kroese, 2014.

[51] I. Boeije, H.R. Bleijenbergh, *Analyseren in kwalitatief onderzoek : denken en doen.* Amsterdam: Boom, derde druk. ed., 2019.

[52] B. J. Oates, *Researching information systems and computing.* Sage, 2005.

[53] P. I. Fusch and L. R. Ness, "Are we there yet? data saturation in qualitative research," *The qualitative report*, vol. 20, no. 9, p. 1408, 2015.

[54] J. Materzynska, G. Berger, I. Bax, and R. Memisevic, "The jester dataset: A large-scale video dataset of human gestures," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 2874–2882, 2019.

[55] O. Kopuklu, "Motion fused frames implementation in pytorch, codes and pretrained models." https://github.com/okankop/MFF-pytorch, Sept. 2021.

[56] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, 2015.

[57] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," *arXiv*, 2021.

[58] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining and knowledge management process*, vol. 5, no. 2, p. 1, 2015.

[59] M. Claesen and B. De Moor, "Hyperparameter search in machine learning," *arXiv*, 2015.

[60] C. L. Devasena, T. Sumathi, V. V. Gomathi, and M. Hemalatha, "Effectiveness evaluation of rule based classifiers for the classification of iris data set," *Bonfring International Journal of Man Machine Interface*, vol. 1, no. Special Issue Inaugural Special Issue, pp. 05–09, 2011.

[61] A. Bulat, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "Toward fast and accurate human pose estimation via soft-gated skip connections," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*, pp. 8–15, 2020.

[62] D. Groos, H. Ramampiaro, and E. A. F. Ihlen, "Efficientpose: Scalable single-person pose estimation," *Applied Intelligence*, vol. 51, no. 4, pp. 2518–2533, 2021.

[63] H. Li, "Learning to rank for information retrieval and natural language processing," *Synthesis lectures on human language technologies*, vol. 7, no. 3, pp. 1–121, 2014.

[64] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.

Appendix A

Interview Question Sign Similarity

This appendix lists the interview questions for the interviews to get insights on how sign language learners perceive similarity between signs used for answering RQ2. This was the first part of the interview.

1 What kind of challenges did you encounter when learning sign language?

2 How did confusing signs play a part?

3 What kind of signs were confusing when learning sign language?

4 How do you deal with/solve those confusing?

5 How would you define similarity in signs?

6 How would you rate position as a source of mistakes/confusability?

7 How would you rate arm configuration as a source of mistakes/confusability?

8 How would you rate hand configuration as a source of mistakes/confusability?

9 Could you rank position, arm/hand configuration as sources of mistakes/confusability?

10 How suitable do you think the given taxonomy is for defining similarity in sign language?

11 When retrieving results, what property of the input sign would like to see back in your results?

12 Do you have any other remarks about mistakes or confusability in signs?

Appendix B

Interview Question Requirements Engineering

This appendix lists the interview questions for the interviews to get insights on the desired functionality and design of a visual dictionary that supports sign language learning used to answer RQ1. This was the second part of the interview.

1 Would you be more interested in using this tool as a dictionary or as a tool to find similar signs?

2 Approximately how many results would you want to retrieve? Would you rather have a single result or a list of multiple options?

3 Would you like to have additional statistics/information on each sign?

4 Retrieving more results will yield more relevant signs, but the ratio of irrelevant signs will be higher. How do you want this trade off handled by the tool?

5 Would you still be willing to use the tool if you have to edit input videos beforehand?

6 Could you think of extra features that would be helpful for sign language learners?

7 Are you aware that your video will be uploaded and processed?

8 Do you have any concerns about this process?

9 Is there any functionality you are missing currently?

10 Do you foresee any problems with the tool?

11 Any other general remarks?