Computational Linguistics in the Netherlands Journal 12 (2022) 287-309 Submitted 12/2022; Published 12/2022

A Sign Similarity Approach to an Information Retrieval Inspired Visual Dictionary for Sign Language Learners

Mark Wijkhuizen[*] MARK.WIJKHUIZEN@GMAIL.COM
Onno Crasborn[**] ONNO.CRASBORN@RU.NL
Martha Larson[*,**] M.LARSON@CS.RU.NL

[*] *Institute for Computing and Information Sciences, Radboud University, Netherlands*

[**] *Center for Language Studies, Radboud University, Netherlands*

Abstract

In this paper, we propose that visual dictionaries for sign language learners should take visual similarity into account in a manner that is inspired by the design and evaluation of information retrieval systems. Sign language learners would benefit from a visual dictionary that allows them to search for the translation of a sign using their web camera to capture themselves executing the sign. As computer vision technology develops towards that goal, we point out that learners are not necessarily supported by systems that return exact matches. Rather, these systems should take sign similarity into account, exhibiting robustness to less-than-perfect sign execution as well as providing information about signs that are visually similar, but different in meaning. The contribution of this paper is a sign similarity measure that was designed based on interviews with signers and sign learners. We also present two sets of experiments to demonstrate how the measure can be used to evaluate a visual dictionary and as an objective function for sign ranking.

1 Introduction

Sign language learners have a variety of dictionaries at their disposal, but these dictionaries currently do not take full advantage of video analysis technology, i.e., computer vision. In this paper, we propose that video analysis should not just be used to recreate the functionalities of existing dictionaries, but should actually be used to develop new possibilities for a visual dictionary. Our focus is on Sign Language of the Netherlands (*Nederlandse Gebarentaal*, abbreviated NGT), but the principles that we propose could be extended to any sign language.

We introduce a sign similarity approach that is suited for use as an NGT visual dictionary. The core of our approach is a *sign similarity measure* that is designed on the basis of interviews with signers and sign learners. Our wider aim is to inspire, by example, machine learning researchers working on sign language to involve signers more directly in the design of machine learning systems for sign language. The sign similarity measure makes it possible to implement and evaluate an information retrieval inspired visual dictionary that takes users' needs directly into account in the ranked list of sign matches that it produces.

The dictionary consists of a large collection of video clips of expert signers producing signs that have been labeled with Dutch translations. Users of the dictionary record a short video of the sign that they are looking for, and the system returns a list of videos of signs that are similar to this sign drawn from the collection. Because the videos in the collection are labeled with Dutch translations, by matching the input signs with the signs in the collection, the dictionary translates from NGT to Dutch.

The new possibilities the we propose that a visual dictionary should offer are illustrated with two examples. First, the dictionary should cover full *variability* of a sign. Figure 1 contains two keyframes from a video of a signer executing the same sign on two different occasions. The examples

Figure 1: Example of two signs that could look visually different to a learner, but in fact should be understood as *the same sign* (Hoogeveen 2020).

Figure 2: Example of two signs that could look visually similar to a learner, but are in fact are *two different signs*. Note the position of the signer's left hand (Hoogeveen 2020).

are drawn from the Universal Declaration of Human Rights (UDHR) in NGT (Hoogeveen 2020), published in the UDHR Translation Project by the United Nations (United Nations n.d.). It can be seen that the exact angle and the facial expression are different. We would like our dictionary to return video clips of both signs to a user searching for a sign, because we posit that seeing the variability of the sign will help sign language learners recognize it more readily in context. Further, we assume that learners are not necessarily able to execute the sign as precisely and correctly as it is executed by an expert signer. We would like our dictionary to be tolerant to the variability in how learners realize signs and be able to match, as well as possible, the videos of learners' attempts with the videos signs in the collection.

Second, the dictionary reveals the *distinguishing characteristics* of a sign. Figure 2 contains two keyframes from a video of a signer executing different signs. It can be seen that the signs differ with respect to the orientation of the left hand. We would like our dictionary to return video clips of both signs to a user searching for a sign to support that user in learning exactly what makes the sign distinct from closely related signs.

The concept for the visual dictionary that we are proposing is a move in the direction of addressing the shortcomings of existing dictionaries. Historically, sign language dictionaries contain words from a spoken language each accompanied by a picture of a person signing the translation in sign language (Zwitserlood 2010). For example, Basic Dictionary of NGT (*Basiswoordenboek NGT*) (Schermer and Koolhof 2018) makes it possible to find the NGT sign translation for Dutch.

Only infrequently do dictionaries translate both directions, i.e., from a spoken language to a sign language *and* from a sign language to a spoken language. The online sign dictionary for NGT (*Gebarenwoordenboek*) (Nederlands Gebarencentrum 2021b) is an example of a dictionary that can translate in both directions. However, to find a sign it is necessary to specify at least one parameter (shape of the left hand, shape of the right hand, hand position, movement, or mouth shape). Our approach aims to reduce the effort needed to use the dictionary by making it possible to search for a sign by creating a short video clip using a webcam. Further, we hope that this input method will support learners who might have seen a sign and who can imitate it approximately, but are not able to correctly articulate all parameters, or learners who are just curious to explore signs similar to one that they have seen.

The concept for our visual dictionary is inspired by *information retrieval*, the technique that underlies search engines, which take a query of input and return a list of relevant results as output. *Relevance* reflects the relationship between input and output that is useful to the user, and is a key notion in information retrieval. The information retrieval framework makes it possible to assess the results in the list according to the degree of relevance. In other words, the framework enables evaluation of the signs returned by our visual dictionary according to varying degrees of similarity, as perceived by signers. Similarity ranges from two signs being identical to two signs being judged by signers as related, but not being exactly identical. In contrast, previous work on visual dictionaries have focuses on results either being correct or not, e.g., Fragkiadakis et al. (2020) and Fragkiadakis and van der Putten (2021), and not on gradations of relevance that could be helpful for the learner. This means that our system evaluation is able to distinguish between a system that is good at returning near matches that are interesting for learners from one that does not.

The main contribution of this work, our sign similarity measure, allows us to evaluate gradations of relevance in lists of signs returned by a visual dictionary. The scheme underlying the similarity measure resembles, an is consistent with, descriptions of the dimensions of signs that are found in the literature, e.g., Stokoe (1980), but is specifically elicited for the use in a language learning application. We propose that visual dictionaries that produce ranked lists of matching signs use such a scheme to evaluate their ability to match an input sign with helpful signs in the sign collection.

Our work is related to a previous study that found that the precision of the list of returned signs impacts users' satisfaction with the results (Alonzo et al. 2019). Like our work, this study focuses on understanding the needs of users of a visual dictionary. The study concludes that users want a list of results that contains an exact match with the sign that they are looking for, but that they also prefer the other signs in the list to be similar to that sign. The difference with our own work is that we produce a sign similarity measure that can be used to assess the usefulness of a list to users. In contrast, Alonzo et al. (2019) assesses the ability of existing metrics to reflect users' needs. We are interested in going beyond the idea that similarity of signs in the results list improves the user's trust in the system to the idea that similarity in the results list actually supports learning.

The paper in organized as follows. In Section 2, we cover the information about related work and the background of our study that is relevant for the paper, including the details of the dataset of web camera videos created for this research. Then, in Section 3, we describe the design of the interviews that we carried out in order to understand sign similarity from the perspective of the intended users of our visual dictionary. Next, in Section 4, we report on the results from the interviews and introduce the similarity measure that we developed on the basis of these results. Section 5 contains additional information from the interviews on general requirements for a visual dictionary, and provides a confirmation of our goal and information about how a visual dictionary could be designed. We present two examples of how the sign similarity measure can be applied in Section 6. First, we demonstrate how it is used for evaluation within an example visual dictionary that is based on Motion Fused Frame and one-shot learning. Then, we demonstrate its potential for use as an objective for learning in a signer ranker that uses learning-to-rank. Finally, Section 7 presents discussion, conclusion, and an outlook onto future work.

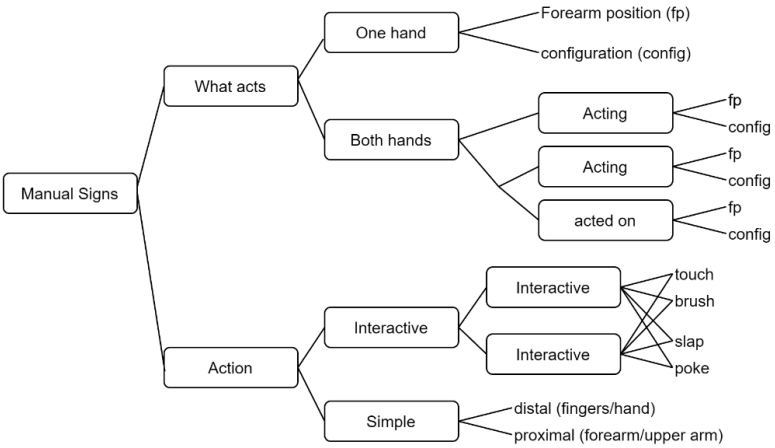

Figure 3: Stokoe's formational division of manual sign language morphemes (Stokoe 1980).

This paper is based on research carried out in the context of a master thesis of the first author (Wijkhuizen 2021), which contains additional information. Selected code helpful for reproducing the paper can be found on github[1].

2 Related work and background

2.1 Learner mistakes and Stokoe's formational division

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 (Bellugi et al. 1974). In spoken languages, it would be more likely that phonetically close words like "cat" and "bat" can be confused. In signed languages, this same phenomenon is present; however, the phonetics are visual, rather than vocal. Signs can be represented in terms of simultaneous independent formational parameters organized along formational dimensions. Independently, the parameters can be realized correctly or incorrectly (Bellugi et al. 1974). For this reason, we can expect there to a systematicity to learner mistakes. It is this systematicity that we hope to capture with our sign similarity measure.

An example, of a formational division of sign language is the one created by Stokoe, shown in fig. 3 (Stokoe 1980). The dimensions of the formational division are orientation (forearm configuration), handshape (hand configuration), location (acted on) and movement (action). A sign can be performed at the correct position, but with a wrong hand shape.

This formational division was originally created for ASL (American Sign Language); however, both ASL and NGT fall under the same French sign language family (Stokoe 1980). For this reason, in this work, we assume that Stokoe's formational division is applicable to NGT as well. Note that sign language also has a nonmanual component. Examples of the non-manual component include mouthing the spoken language equivalent of a word without producing the actual sound as well as oral components where the articulation does not imitate an existing word (Nederlands Gebarencentrum 2021a). The nonmanual component is not included in Stokoe's formational division.

The formational division helps to illustrate the key difference between mistakes made in spoken and signed languages. Sign languages are parallel, in the sense that the building blocks of the sign are realized simultaneously, whereas the building blocks of words in spoken language of ordered (Bellugi

1. `https://github.com/MarkWijkhuizen/Supporting-Sign-Language-Learning-With-a-Visual-Dictionary`

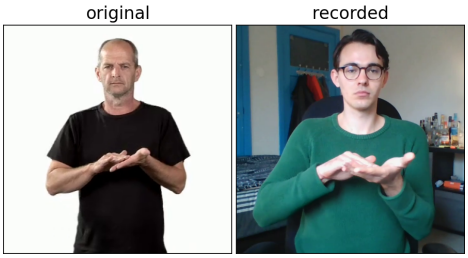

Figure 4: Studio frame (background collection/training collection) vs. self recorded frame (validation and test) of the sign for *zalf* (ointment)

and Fischer 1972). This parallel nature of signs requires special attention in 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 sign similarity that are used to create our sign similarity measure. Note that despite the differences the building blocks of sign and spoken languages, in both cases, these building blocks are referred to as *phonemes*. In the literature, *phonemes* and *morphemes* are often used interchangeably to designate the aspects of a sign. Here, we will us the designation *phonemes*, to reflect that the basic building blocks of sign cannot, in general, be teased apart from the signs that they compose without alterning the meaning or function of the sign.

2.2 Our NGT datasets

In this work, we use two datasets, which are described in this section. The first set, the training set, is the background collection of labeled data that is used to implement that visual dictionary. It is called the "training set" because it is used to train the visual dictionaries in Section 6. The training set consists of 300 video clips of 300 different NGT signs. The clips are randomly drawn from the *Nederlandse Gebarentaal* dataset of Global Signbank (Global Signbank n.d.). The videos are annotated with a conventional inventory of phonemes related to how signs are articulated. The phonemes express, for example, location and direction of movement. The training videos are recorded under studio conditions with an experienced signer, which results in videos of high quality, with flawless articulation and no background noise.

The second dataset was created specifically for the purpose of this research in order to be representative of video clips recorded by sign language learners using their own web cameras. An illustration of the contrast of a studio video in the first dataset and a learner web camera video in the dataset we created is provided in figure 4.

It was necessary to create a specific dataset because our visual dictionary is intended to be used by language learners recording themselves using their web cameras. The signs of the users using the visual dictionary will likely not be articulated flawlessly, since sign language learners can make articulation errors or will be using the dictionary to look up signs which they do not remember correctly. To obtain a representative validation and test, set six volunteers, including the first author, recorded 253 videos of signs in home environments using the built-in laptop webcam or external webcam. The 253 signs are a random subset of the 300 signs used for training the system and are assigned the same sign and phoneme labels as the training recordings. 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 1.

Instructions were given to help ensure that the data collected would reflect real-world conditions under which a visual dictionary would be used by sign language learners. Participants viewed each sign and and were asked practice its articulation before recording the sign. Practice was deemed necessary to ensure a minimal articulation level by participants with no signing experience. Five

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

out of the six volunteers had no signing experience and one was an experienced signer. In addition, the recording environments were not cleaned up, but rather were cluttered backgrounds typical of home environments were captured. Specifically, the recording locations were the places at which the volunteers usually used their computers. Volunteers generally used a conventional desk as their recording location. Two exceptions were a participant using a standing desk and a participant using a couch. In some cases, the office chair was located slightly further back than its usual position to capture the whole signing space in the recording. Chair position was the only adaptation made in comparison to the natural computer usage setting of the participants. Each sign in the validation and test data is labeled with the name of the sign. This label is used to evaluate the accuracy of the visual dictionary. Each sign is also labeled with the same phoneme labels as the corresponding sign in the training data.

2.3 Visual dictionaries and Motion Fused Frames (MFF)

Computer vision has great promise for visual dictionaries over the years. Over ten years ago, DTW (Dynamic Time Warping), a method for matching sequences of potentially unequal lengths, was successfully been applied to a dataset of 1113 signs with a $top10Acc$ of 78% (Wang et al. 2010). This system did, however, require users to manually mark the start and end of the sign in the video and 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 generalizability of the results to real world setting is questionable. A deep learning approach without human assistance would streamline the information retrieval process. Recently, Fragkiadakis and van der Putten (2021) 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.

The sign similarity approach that we propose in this paper is applicable to any search-based visual dictionary. To demonstrate it in action, we implement a basic visual dictionary and use our sign similarity measure to evaluate its performance. We also implement a second visual dictionary that integrates our sign similarity measure directly in its design, using it as an objective function.

The full range of ways that signs can be captured has been discussed in previous work: RGB (Red, Green, Blue), depth camera, or thermal camera (Rastgoo et al. 2021). RGB videos can be further processed to create skeletal videos to solely represent body movement and posture (Rastgoo et al. 2021). These input modalities can be combined to improve the performance of these models, which makes them more complex and harder to train (Elboushaki et al. 2020) (Min et al. 2020). However, despite the different approaches, algorithms for sign search tend to focus on the spatial properties of an action, not the temporal.

In our work, we integrate motion by taking advantage of Optical Flow (OF) frames, which represent the motion ongoing at a moment in a video as a single frame. OF is the estimated displacement field between two consecutive frames and OF frames capture the horizontal and vertical movement on a pixel level (Brox et al. 2004). We adopt the idea of Kopuklu et al. (2018), who

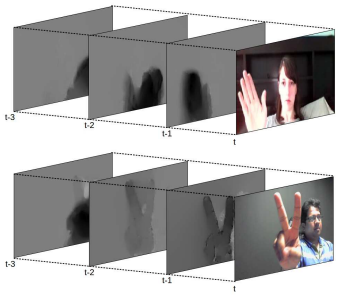

Figure 5: Motion Fused Frames: Optical Flow frames are appended to RGB frames (Kopuklu et al. 2018)

propose a data fusion method where Optical Flow frames are appended to RGB frames, thereby fusing temporal and spatial information to create a single image called an MFF (Motion Fused Frame).

An MFF captures both the current state in the RGB frames, i.e., the location and hand shape, as well as the movement between states. A schematic representation of an MFF is shown in Figure 5. The OF frames of an MFF represent the states previous to a given RGB frame. Kopuklu et al. (2018) apply MFFs in the context of conventional sign language classification. Alone, RGB frames fall short of capturing the complete structure of a sign, since they exclude movement. In our work, we apply MFFs in one-shot classification for the purpose of creating the ranked list of signs returned by the visual dictionary.

An advantage of MFFs is that they require only one model to be trained, in contrast to decision or feature level fusion which requires one model for each data modality (Elboushaki et al. 2020). In our work, as explained in Section 6, MFFs are the input to the neural network that is used for one-shot learning.

3 Design of interviews

We carried out a set of interviews to establish the requirements for our sign similarity approach to an information-retrieval-based dictionary for sign language learners. The interviews helped to clarify the functionality important for sign language learners and what learners would like the user interface to look like. Most importantly, the interviews allowed us to understand how sign similarity is perceived by sign language learners, which allowed us to design our sign-similarity measure. In this section, we present the design of the interviews in detail, to make it possible for future research to reproduce our interview setting and confirm our findings.

3.1 Participants

The group of interview participants consisted of a sign language teacher, a PhD student on sign language classification, a sign language interpreter and five sign language students of various levels. The interviewees' backgrounds are listed in Table 2.

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, and was interviewed via an interpreter. The PhD candidate is not an NGT signer, but can sign in International Sign and American Sign Language (ASL). Lastly, the levels of the NGT students varied from a semester long minor to a fourth year student.

Interviewee	Sign language background	Sign Languages
1A	NGT Interpreter	NGT, ASL
1B	NGT teacher	NGT
2	4th year student NGT interpreter	NGT
3	PhD candidate in 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 2: Interview participants

3.2 Opening and key principles

Before the interview started a few minutes were reserved to make the interviewee feel comfortable. The interviewer (the first author) introduced himself and asked some general questions so that he and the interviewee could get to know each other. Next, the interviewer introduced the research, mentioned the duration of the interview, and discussed informed consent. A PDF consent form had been sent to the participants before the interview. If the interviewee had not filled in the consent form beforehand it was filled in during the discussion at the start of the interview. The first question of the interview was a warm-up question to introduce the topic and give the interviewee time to bring up any memories about the general topic.

Interview questions are carefully constructed to have a neutral tone. During the interview the interviewer was careful to respond neutrally to answers, both verbally and non-verbally. Any follow up questions were kept neutral as well to avoid implying answers. The follow up questions aimed to make interviewees think more carefully about a certain topic to get a more comprehensive answer. Each part of the interview ends with the interviewer asking the interviewee if there are any other general points that came to mind about the topic. This measure was intended to give space for the the interviewee to talk freely about the topic to discuss anything that was not covered by a specific question.

3.3 Sign similarity

The central question to be answered during this part of the interview was how sign language learners perceive similarity between signs. This information was necessarily in order for us to define a sign similarity measure that could capture the relevance of the results returned by a visual dictionary for sign language learners. This part of the interview was designed using a funneling approach, starting with broad questions on confusion signs and mistakes during signing and thereby implicitly on sign similarity. A full list of questions used during the sign similarity interview is available in Appendix A.

We used a direct question on how to define sign similarity, which was used to elicit what the first thought that came to the participants' minds 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 in the results of a visual dictionary. The interview was finalised with an open question to discuss anything that comes up regarding sign similarity.

3.4 General requirements

The second part of the interviews concerned the requirements potential users have for the system. A full list of questions used during the sign similarity interview is available in the original master

thesis (Wijkhuizen 2021). Questions mainly concerned additional functionality, in addition to retrieving videos of signs, that would help to support sign language learning. The questions regarding functionality lead to 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.

4 Sign similarity interview results

In this section, we present the results of the sign similarity and describe how we built on these results to create our sign similarity measure.

4.1 General results

We start with describing the inductive coding process that was used to analyze the interviews and introducing the overall results. In the first step, we read through the interviews and identified open codes corresponding to broad topics in individual quotes from the interviewees. Next, we identified axial codes by grouping quotes within the open codes. Finally, we determined selective codes that express the specific insight that was obtained from the interview quotes associated with a given axial code.

The results of the interviews are shown in Table 3. The broad topics are *location*, *nonmanual*, *movement* and *handshape*, and are shown as gray bars in Table 3. A fifth code, *other* was assigned to all remaining interview quotes. The axial codes are shown on the left and the corresponding specific code on the right.

The broad topics that emerged from the interview form the basis for our sign similarity measure. Within the broad topics, we look for the dimensions of sign language that are relevant to language learners. We identify *location*, *movement* and *handshape* as essential. Non-manuals are an integral part of sign language, however, the interviewees did not expect them to be used by sign language learners in a visual dictionary. We also add *handedness*, which is an axial code of *handshape*, but is found by interviewees to be important for restricting results.

In the rest of this section we will discuss the insight that the interviewees provided on each dimension further and describe how we integrated it to obtain our sign similarity measure. The next four subsections, Section 4.2 on *location*, Section 4.3 on *movement*, Section 4.4 on *handshape*, Section 4.5 on *handedness*, discuss the four dimensions. Each subsection starts with additional information gained from the interviews and then described how we used this information to create a measure for the dimension. The basic process we use to create a measure is to used the information from the interview to organize the phonemes with which the training data is annotated into a tree structure. The tree structure is created so phonemes that sign language learners find confusable are located close to each other in the tree, for example, they are sister leaves. As a result, distance calculated within the tree reflects the type of similarity between signs that is important for sign language learners. The final subsection, Section 4.6 discusses how the measures for the dimensions are combined to create the overall sign similarity measure.

4.2 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.

Location	
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
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 3: Open, axial and selective inductive codes of the interview on sign language similarity. Quotes are firstly clustered in open codes which are denoted in gray 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.

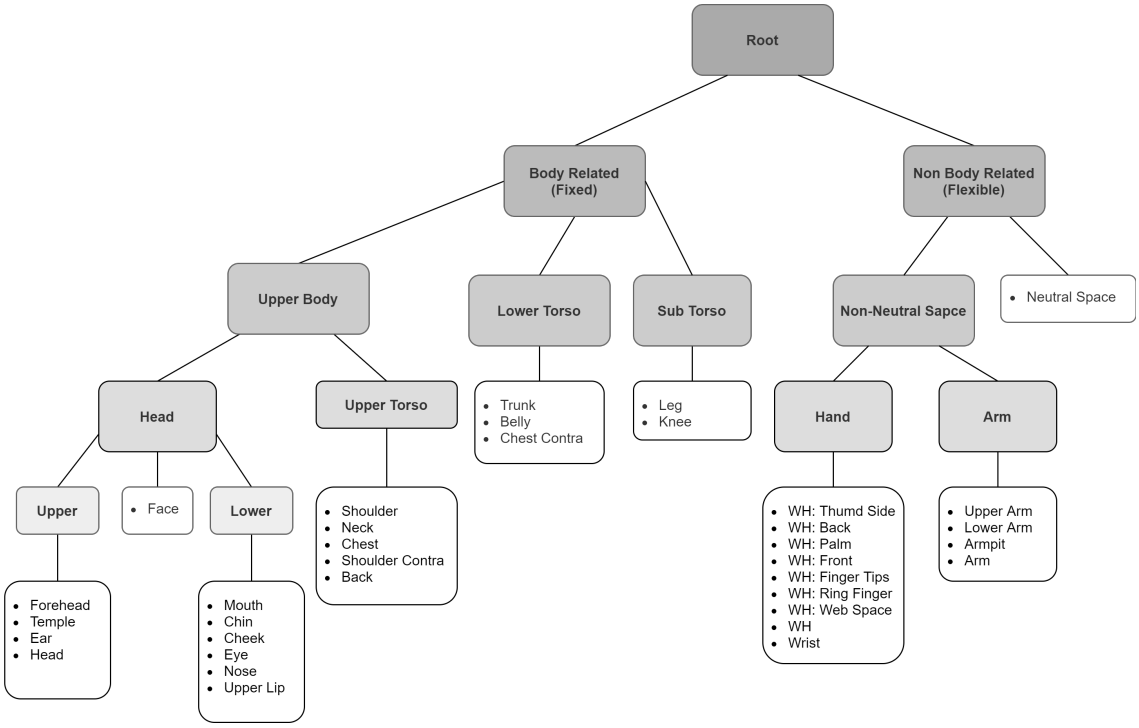

Figure 6: Sign location taxonomy tree where leaves are itemised for readability purposes. WH stands for Weak Hand. The locations within a region are easily confused with each other.

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 so 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 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, shown in Figure 6. Leaves are listed for readability.

Similarity quantification is computed by computing the node distance between leaves as fraction of the longest distance and subtracting it from 1. The result is a similarity score in the practical range $[0, 1]$. The node distance is the number of nodes that must be traversed to get from one location to another location within the tree. It is equal to the number of edges between two locations. The node distance is subtracted from one so that a longer node distance results in a lower similarity score.

The longest node distance is 9, 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 calculated on the path $forehead \Rightarrow upper \Rightarrow head \Rightarrow upper\ body \Rightarrow upper\ torso \Rightarrow shoulder$, which are five steps. Similarity between forehead and ear is $1 - \frac{2}{9} \approx 0.78$. Note that the listed leaves should be interpreted as single leaves, making the path between forehead and ear $forehead \Rightarrow upper \Rightarrow ear$.

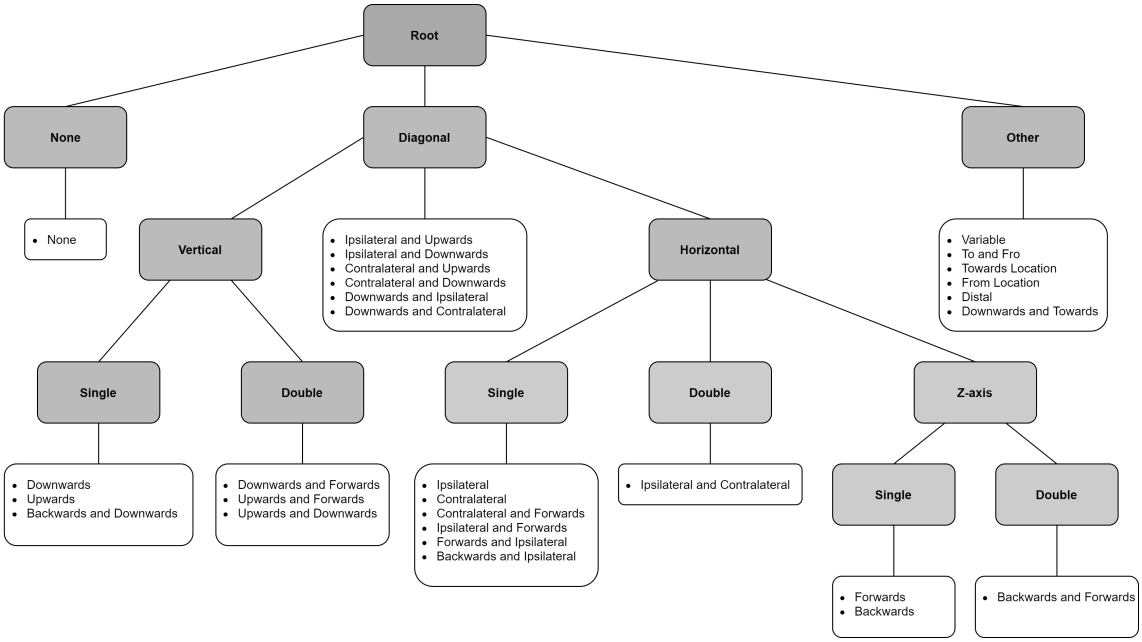

Figure 7: Sign movement taxonomy tree where leaves are itemised for readability purposes

4.3 Movement

When discussing similarity in movement, the interview 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 dimension. 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.

Several design choices have been made to best structure the distances. 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 $1 - \frac{|single,z\text{-}axis,horizontal,single,ipsilateral|}{7} = \frac{5}{7} \approx 0.29$. This structure can convert any pair of movements directions to a similarity value in the range [0, 1], based on visual axial similarity.

4.4 Handshape

When discussing similarity in handshape during the interviews, participants mentioned there were many similar looking handshapes, 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 work of Van der Kooij (2002) on phonological categories of sign language in NGT provided a mapping from handshape to articulator properties. 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 (Van der Kooij 2002). 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.

4.5 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 dimension in Stokoe's formational division of sign language (Stokoe 1980). 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 as well as on the interviews.

The handedness similarity is, firstly, divided into one-handed and two-handed signs. The latter is sub-divided into different configuration. The result is the taxonomy tree shown in Figure 8. Distance between handedness configuration is measured as node distance proportional to the maximum

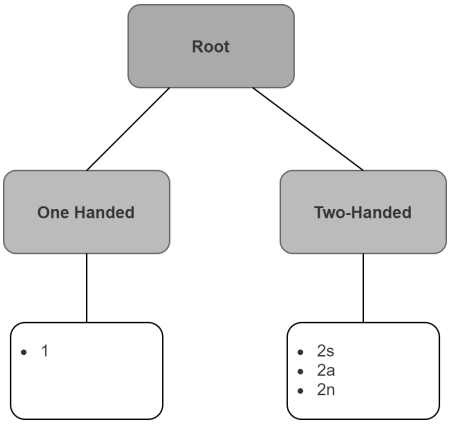

Figure 8: Handedness taxonomy tree where leaves are itemised for readability purposes

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 $1 - \frac{2}{4} = 0.50$.

4.6 Overall sign similarity measure

The four dimensions location, movement, handshape and handedness are assembled in order to quantify sign similarity. Handshape is subdivided into strong hand and weak hand, resulting in five similarity scores. 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 (Radlinski and Craswell 2010) to measure the relevance of the retrieved signs. The NDCG formula is shown in eq. (1). Here, Q denotes the number of queries, N_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 $rel(d^i j)$ the relevance of the j^{th} retrieved document with respect to query q_i.

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

The relevance of a document is given by $\frac{2^{rel(d^i j)}-1}{log(j+1)}$, 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.

5 General requirements interview results

In this section, we summarize the results of the general requirements part of the interviews. These results were not used to design the sign similarity measure, but are nonetheless interesting for future work on visual dictionaries.

Applying the inductive coding process on the general interviews gave rise to four themes: *use case*, *result layout*, *additional information* and *video processing*. These are shown as the gray rows

in Table 4. For each theme, the axial codes are provided below, along with a user story, which spells out the functionality that users desire to see in a system.

The main results relevant for this paper are related to the first open code, which concerns the *use case* of the visual dictionary. The interviews revealed that users primarily anticipated that they would use the system as a dictionary, not as a tool to look up similar signs. However, users would like to have the ability to see semantically similar signs, when they are looking for a sign. This functionality would help them to extend their vocabulary in the context they are working in. The interviewees anticipated that sign language learners would not articulate properly and would need to view multiple results to find the sign that they were looking for. In sum, it seems that users of the dictionary would need the coverage of variability, illustrated by Figure 1 above. They would also possibly appreciate similar signs that would reveal distinguishing characteristics, illustrated by Fig. 2 above. However, we need to make sure that we do not overestimate the importance of distinguishing characteristics, since the interviewees did not feel they would actually want to go in search of similar signs.

The other open codes are interesting for design of future search-based visual dictionaries. We cover them here, but only briefly, since our main focus is on the sign similarity approach. For more details, see the original thesis (Wijkhuizen 2021). Interviewees mentioned several points regarding the *result layout*. Varying the desired number of results from 6 to 20, where 10 seemed an appropriate balance between convenience and completeness. They mention offering less results in the case of low confidence and keeping the interface simple. *Additional information* was also important and included a desire for explanation of iconicity, usage, and also a view of the sign from the side. Finally, the interviewees mentioned *video processing*, indicating a preference for a video trimmer built into the tool and a wish for privacy protection.

6 Validation Experiments

In this section, we present experiments on two visual dictionaries that demonstrate the use of our sign similarity measure. Sections 6.1and 6.2 report on a one-shot learning approach to a visual dictionary based on Motion Fused Frames (MFFs). These experiments demonstrate the sign similarity measure being applied to evaluate a visual dictionary. Section 6.3 report on a Learning-To-Rank (LTR) approach also based on MFFs. This experiment demonstrate the sign similarity measure being used directly in the machine learning algorithm that is used by the visual dictionary.

6.1 Evaluating a visual dictionary with our sign similarity measure

6.1.1 One-shot learning visual dictionary based on MFFs

First, we describe the machine learning algorithm that we use to implement our visual dictionary for the purpose of our validation experiment. The goal of this experiment is to demonstrate results reported using our similarity measure. Recall from Section 1 that the similarity measure could be used to assess any visual dictionary that produces a ranked list of results.

Our visual dictionary is a one-shot classifier trained on Motion Fused Frames (MFFs), which were described in Section 2.3. The classifier is trained to recognize the 300 signs in the training data. We need a one-shot classifier because each sign is only included once in the training data.

When users look a sign up in the dictionary, they record a video, which is converted to MFFs and passed to the classifier. The classifier produces a list of signs that it determines are probably the sign in the video. The signs are ranked in order of the probability produced by the classifier. The dictionary returns the videos from the training data corresponding to the top-ranked signs.

Here, we describe the machine learning algorithm in more detail. An MFF representation can be configured with varying number of MFFs, optical flow frames, and color frames. Increasing the number of MFFs results in more spatial information being captured, whereas increasing the

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 4: 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.

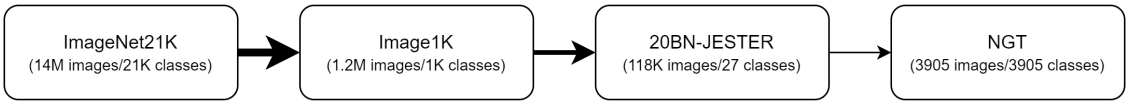

Figure 9: Transfer learning process from the huge general ImageNet21K dataset to one-shot learning NGT dataset. The first two steps generic features are learned, whereas the third step learns features close to the sign language domain.

number of Optical Flow frames results in more temporal information being captured. This video representation can transform a video of arbitrary length to a fixed number of frames which is typically required for machine learning models (Elboushaki et al. 2020)(Min et al. 2020)(Kopuklu et al. 2018). This video representation has been successfully applied to the Jester dataset (Kopuklu et al. 2018). This dataset contains 148,092 videos of gestures performed by 1300 volunteers recorded in front of a webcam in an unconstrained environment (Materzynska et al. 2019). 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.

The classifier is convolutional neural network (CNN), specifically, the EfficientV2-S model. It achieves a top-1 accuracy of 84.9% on the ImageNet classification task (Tan and Le 2021). The CNN architecture is created using a neural architecture search which is set to optimise accuracy, parameter efficiency and training efficiency. The experiments were carried out on a laptop with a gtx1060 GPU and results were returned in less than a second. All data preprocessing and programming were done locally, only training was done on a server (NVidia V100).

One-shot-learning 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} (Wang et al. 2020). 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 randomly 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 that 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 9. The randomly initialised parameters are first fitted to the huge ImageNet21K training set to learn generic features. Next, the model is fine tuned on the large ImageNet1K dataset. Although ImageNet21K and ImageNet1K are very large datasets, they contain everyday images of, e.g., animals and vegetables, which are visually 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.

The Jester dataset containing gesture videos 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 (Wang et al. 2020). 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.

6.2 Visual Dictionary Results

Table 5 provides the results of the evaluation of our one-shot classifier on the validation set and the test set in terms of the conventional evaluation metric used for visual dictionaries (TopK accuracy)

MFF Config.	Val top20Acc.	Val NDCG@20	Test Acc@20	Test NDCG@20
RGB only	0.46	0.51	0.36	0.55
RGB + 3 OF frames	0.60	0.59	0.44	0.58

Table 5: Results of the visual dictionary with an without using Optical Flow frames (OF frames) on the validation and test set. Accuracy (Acc) is the conventional evaluation metric and NDCG represents the similarity measure that is introduced in this paper.

	NDCG@20				
	Location	**Movement**	**Strong Hand**	**Weak Hand**	**Handedness**
Validation	0.63	0.39	0.29	0.28	0.50
Test	0.64	0.37	0.26	0.29	0.50
Baseline	0.585	0.354	0.227	0.241	0.453

Table 6: Phoneme $NDCG$@20 for each on both the validation and test set and the corresponding baseline scores

and the evaluation metric proposed in this paper (NDCG based on the similarity measure). TopK accuracy is defined as,

$$topKAccuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + False\ Positives + True\ Negatives + False\ Negatives} \quad (2)$$

and NDCG is defined as in 1 in Section 4.6.

The first row shows the performance achieved using only RGB information, which does not represent motion. The second row shows the performance achieved when three Optical Flow frames are added. Our experiments determine that three was a good number of Optical Flow frames to add, but interestingly that there was no real difference in adding one, two or three frames. Additional experiments illustrating this point are included in Appendix B.

The main message of Table 5 relevant for the sign similarity approach (i.e., the NDCG) introduced in this paper is straightforward. We see that the NDCG and Accuracy are correlated, reflecting that the sign similarity approach is picking up on the exact matches that are returned by the dictionary. Second, we notice that although they behave similarly, they are not completely synchronized. Specifically, adding Optical Flow frames seems to improve the accuracy on both the validation and test set more than it improves the NDCG. Further investigation is necessary to further validate what our sign similarity NDCG is capturing. However, Table 5 provides some basic evidence that if system development focuses only on maximizing a metric like accuracy, it might be missing aspects that are picked up by sign similarity NDCG, and are important to language learners using the dictionary.

In order to gain more insight into what our visual dictionary is capturing, the NDCG score per dimension is analysed. We report a baseline computed using 1000 random ranking of the complete 300 training samples. The result are shown in Table 6.

Both the validation and test $NDCG$@20 score consistently better on all dimensions. There is no clear dominant dimension the model captures best. Given that MFFs use Optical Flow frames one might expect movement to be better modeled; however, this is not the case. Strong and weak hand shapes are surprisingly well picked up, which was not expected given the low video quality.

	topKaccuracy (%)				NDCG (%)			
	@1	**@5**	**@10**	**@20**	**@1**	**@5**	**@10**	**@20**
RGB + 3 OF frames	0.6	0.21	0.36	0.44	0.63	0.57	0.57	0.58
L2R (with DLS and KFs)	0.4	0.16	0.22	0.36	0.66	0.68	0.69	0.72

Table 7: Performance comparison on the test data between the original one-shot visual dictionary RGB + 3 OF frames and a learning-to-rank (LTR) visual dictionary with Dynamic Loss Scaling (DLS) and keyframe (KF) selection

6.3 Sign similarity as an objective

In this section, we investigate the suitability of our measure for use as an objective function. We recast our visual dictionary as a Learning-to-Rank system (Li 2014) that uses a pointwise approach. The learning to rank system optimizes directly for NDCG calculated on the basis of the sign similarity score produced by our sign similarity measure. Specifically, given an input sign X in the training data, represented as a series of Motion Fused Frames, the learning objective is to predict the NDCG relevance score y (in the range $[0, 1]$) for each other sign in the training set. We used 200 to 500 epochs, dropout 0.25% and a learning rate of 0.001.

We introduce two improvements to the basic LTR system. First, we implemented a keyframe (KF) selection method based on the body keypoint detection tool OpenPose (Cao et al. 2020). The method identifies the most discriminative parts of each sign, and uses only MFFs extracted at these points. By selecting discriminative keyframes, we help to ensure that the MFFs of the users video and the training set videos are well aligned. Full details of the keyframe selection method are available in Wijkhuizen (2021). Second, we use binary cross-entropy loss with Dynamic Loss Scaling (DLS), which scales the loss 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 too large. DLS scales the loss of signs with respect to the relevance, where a higher relevance results in larger adjustment of the loss.

Results are shown in Table 7. As expected the L2R approach which is optimized with respect to our similarity measure performs better with respect to NDCG, which is based on the similarity measure. This demonstrates the basic suitability of the sign similarity score to be used in an objective function. Note that we do not claim that L2R improves over the original one-shot approach. It depends on which metric is considered. L2R lags with respect to accuracy. For completeness we note that the accuracy gap can be closed by also applying Target Loss Scaling in the L2R approach, which is described in detail in Wijkhuizen (2021). Further work is needed to understand in detail the degree to which the NDCG based on our sign similarity score is able to better match the usefulness of visual dictionary results for sign language learners.

7 Discussion and Conclusion

In this paper, we have introduced a sign similarity approach to evaluating visual dictionaries that is inspired by the concept of relevance that is important in information retrieval. The approach is based on a set of interviews with the target group for the visual dictionaries, namely signers and learners of sign language.

The interviews allowed us to identify four dimensions, which taken together define sign similarity. These dimensions are *location*, *movement*, *handshape* and *handedness*. 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. 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 *handshape* could not be determined during interviews and articulator properties based on Van der Kooij (2002) are used instead. 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]. In addition to providing the basis for our similarity measure, these findings also give a comprehensive analysis on how sign language learners perceive similarity between signs.

We have implemented a visual dictionary using Motion Fused Frames, to capture movement, and a one-shot learning classifier. We presented some example results using this system which were evaluated using NDCG, where the grades of relevance were derived from our similarity measure. We have also implemented a visual dictionary using Motion Fused Frames and Learning-to-Rank. The purpose of these experiment was to provide a very basic example of the similarity measure at work, both as an means for evaluating a visual dictionary and also as an objective function for training a visual dictionary approach.

The goal of this paper has been to show the way forward to pursuing a similarity based approach to developing visual dictionary that is based in the needs of sign language learners. We hope that it is a step towards integrating research in the computer vision field and research in the area of sign language. Moving forward, our sign similarity metric should be validated in future user studies, such as those carried out by Alonzo et al. (2019). As mentioned above, validation would provide more information about the degree to which our sign similarity measure succeeds in reflecting the relevance of visual dictionary results for sign language learners. Further, it is important to test visual dictionaries using a more sophisticated ground truth. In our experiments, the phoneme labels of the signs in the validation and test sets do not take inaccuracy of learners into account because they are adopted from the corresponding signs in the training data. Future work should create validation and test sets that are annotated with the literal phonemic labels of what the learner is actually signing and not the ideal labels. Such ground truth will make possible more detailed insights into the ability of our sign similarity measure to support sign language learners.

As visual dictionaries continue to evolve, we urge developers to take into consideration that the input sign to a visual dictionary might be made by a beginning learner of sign language, who can execute an approximation of the sign that they are looking for, but does not have the memory and/or experience to necessarily sign completely correctly. We also find it important not to assume that the user of the dictionary is only interested in one sign and that there is a single best video to represent this sign. Instead, we find that language learners may be supported by also seeing signs that are visually similar, but different from the sign that they are looking for.

8 Acknowledgements

We would like to thank all the participants who took part in the interviews and the volunteers who created the validation and test sets of web camera videos. Also, thank you to Javier Martínez Rodríguez for discussion and support.

References

Alonzo, Oliver, Abraham Glasser, and Matt Huenerfauth (2019), Effect of automatic sign recognition performance on the usability of video-based search interfaces for sign language dictionaries, *The 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*, pp. 56–67.

Bellugi, Ursula and Susan Fischer (1972), A comparison of sign language and spoken language, *Cognition* **1** (2-3), pp. 173–200, Elsevier.

Bellugi, Ursula, Edward S. Klima, and Patricia Siple (1974), Remembering in signs, *Cognition* **3** (2), pp. 93–125, Elsevier.

Brox, Thomas, Andrés Bruhn, Nils Papenberg, and Joachim Weickert (2004), High accuracy optical flow estimation based on a theory for warping, *European conference on computer vision (ECCV)*, pp. 25–36.

Cao, Zhe, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh (2020), OpenPose: Realtime multi-person 2D pose estimation using part affinity fields, *IEEE transactions on pattern analysis and machine intelligence* **43** (1), pp. 172–186.

Elboushaki, A., R. Hannane, K. Afdel, and L. Koutti (2020), 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* **139**, pp. 112829.

Fragkiadakis, Manolis and Peter van der Putten (2021), Sign and search: Sign search functionality for sign language lexica, *The 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pp. 23–32.

Fragkiadakis, Manolis, Victoria Nyst, and Peter van der Putten (2020), Signing as input for a dictionary query: Matching signs based on joint positions of the dominant hand, *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.

Global Signbank (n.d.), Nederlandse gebarentaal dataset. Accessed: 5 December 2022. `https://signbank.cls.ru.nl/datasets/NGT`.

Hoogeveen, Dennis (2020), Universal declaration of human rights in Sign Language of the Netherlands. Accessed: 5 December 2022. `https://youtu.be/lEfNMZaOKFk`.

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

Li, Hang (2014), Learning to rank for information retrieval and natural language processing, *Synthesis lectures on human language technologies* **7** (3), pp. 1–121, Morgan & Claypool Publishers.

Materzynska, J., G. Berger, I. Bax, and R. Memisevic (2019), The Jester Dataset: A large-scale video dataset of human gestures, *2019 IEEE/CVF International Conference on Computer Vision Workshops*, pp. 2874–2882.

Min, Y., Y. Zhang, X. Chen, and X Chai (2020), An efficient PointLSTM for point clouds based gesture recognition, *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5760–5769.

Nederlands Gebarencentrum (2021a), De gesproken component in NGT. Accessed: 5 December 2022. https://www.gebarencentrum.nl/gc.

Nederlands Gebarencentrum (2021b), Online Gebarenwoordenboek. Accessed: 5 December 2022. https://ow.gebarencentrum.nl/search?m=parameters.

Radlinski, Filip and Nick Craswell (2010), Comparing the sensitivity of information retrieval metrics, *The 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 667–674.

Rastgoo, Razieh, Kourosh Kiani, and Sergio Escalera (2021), Sign language recognition: A deep survey, *Expert Systems with Applications* **164**, pp. 113794.

Schermer, Trude and Corline. Koolhof (2018), *Van Dale basiswoordenboek Nederlandse gebarentaal*, eerste editie, zesde oplage. ed., Nederlands Gebarencentrum, Bunnik.

Stokoe, W. C. (1980), Sign language structure, *Annual Review of Anthropology* **9** (1), pp. 365–390.

Tan, Mingxing and Quoc Le (2021), EfficientNetV2: Smaller models and faster training, *in* Meila, Marina and Tong Zhang, editors, *The 38th International Conference on Machine Learning*, Vol. 139 of *Proceedings of Machine Learning Research (PMLR)*, pp. 10096–10106.

United Nations (n.d.), UDHR in sign languages. Accessed: 5 December 2022. https://www.ohchr.org/en/human-rights/universal-declaration/udhr-sign-languages.

Van der Kooij, Els (2002), Phonological categories in Sign Language of the Netherlands, *The Role of Phonetic Implementation and Iconicity. LOT, Utrecht.*

Wang, Haijing, Alexandra Stefan, Sajjad Moradi, Vassilis Athitsos, Carol Neidle, and Farhad Kamangar (2010), A system for large vocabulary sign search, *European Conference on Computer Vision (ECCV)*, pp. 342–353.

Wang, Yaqing, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020), Generalizing from a few examples: A survey on few-shot learning, *ACM Computing Surveys (CSUR)* **53** (3), pp. 1–34.

Wijkhuizen, Mark (2021), *Supporting sign language learning with a visual dictionary*, Master's thesis, Radboud University. https://www.ru.nl/publish/pages/769526/mark_wijkhuizen.pdf.

Zwitserlood, Inge (2010), Sign language lexicography in the early 21st century and a recently published dictionary of sign language of the netherlands, *International Journal of Lexicography* **23** (4), pp. 443–476.

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.

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?

B Additional MFF results

To measure the added value of appended optical flow frames to RGB frames the performance of different MFF configurations is measured. In each step, we increase the number of optical flow frames (OF frames). The number of RGB frames has not been modified from the original paper. The result are shown below in table 8.

Performance/MFF Config.	**0 OF frames**	**1 OF fram**	**2 OF frames**	**3 OF frames**
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: 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 interesting for several reasons. First, the added value of the first optical flow frame is high with a validation *top*1*Acc.* 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, neither for the Jesture data set, nor for the NGT Validation or Test sets.