Digital tools for sign language research: towards recognition and comparison of lexical signs

Fragkiadakis, M.

CHAPTER 3

Signing as Input for a Dictionary Query: Matching Signs Based on Joint Positions of the Dominant Hand

This chapter presents a novel method for searching sign language dictionaries by using a full sign as input. Users can look up signs by performing them in front of a webcam. The method compares the recorded sign with those in the lexicon. This approach, which complements traditional gloss-based or phonological feature searches, utilizes OpenPose to analyze body and finger joint positions. Comparison with lexicon signs is achieved by quantifying and comparing variations in trajectories of the dominant hand and fingers using Dynamic Time Warping (DTW). Tested on ten subjects with varying sign language proficiency, the method was applied to the Ghanaian Sign Language lexicon. Participants replicated 20 signs, resulting in 87% and 74% accuracy for Top-10 and Top-5 rankings, respectively, using only the dominant hand's trajectory. Higher proficiency signers reached 90% accuracy at the Top-10 level. [1]

[1] Chapter based on: Manolis Fragkiadakis, Victoria Nyst, and Peter van der Putten. "Signing as input for a dictionary query: matching signs based on joint positions of the dominant hand". English. In: Proceedings of the LREC2020 9th Workshop on

3.1 Introduction

In most sign language dictionaries, users can search a sign through a written gloss, a unique identifier that by definition refers to a sign. In some cases, the lexica offer the possibility to specify formal parameters of the target sign, for instance, its handshape and location (figure 3.1). The Flemish Sign Language (VGT) dictionary [163], the Swedish Sign Language [81] and the Danish Sign Language [30] are some examples of such dictionaries. After the input, the user is offered a set of signs that match the selected properties which can be then viewed individually.

Although sign search functionality on the basis of a sign parameter value is a useful attribute of sign language lexica, dictionary compilers still have to link these values to the videos. Also, as Zwitserlood discusses, the users of such dictionaries must "abstract away from the sign as a whole" if they want to use the parameter search functionality [184]. Even then, only signs that match the query 100% are returned, and there is no concept of an ordered set of results that match to some degree. A thorough overview of sign language lexica and their features can be found in Zwitserlood's review (2010).

In this chapter we describe our "find the sign" methodology that allows inputting a full video-recorded sign to search for entries in a dictionary. This method requires no training of any kind of model such as the ones used for sign language recognition tasks. In its core, it is a comparison method to quantify the difference in the movement between signs. As a result, it can be used for any sign language. By utilizing a pre-trained pose estimation framework we extract the body and hand joint positions from users using their webcam. Subsequently, by employing Dynamic Time Warping we find the closest matching signs from a compiled lexicon. To date, this methodology has only been applied to sign language classification tasks [84, 144, 158] and not as a mode to complement sign search possibly solving the problem of ordering retrieval previously discussed. Additionally, we have developed a visualization tool to allow researchers to view the rendered paths of the dominant hand to further explore the overall difference in signing movements.

---

the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives. Marseille, France: European Language Resources Association (ELRA), 2020, pp. 69–74.

The chapter is structured as follows: in Section 3.2 we give an overview of methods that utilize Dynamic Time Warping in the gestural and sign language domain. In Section 3.3 we describe our methodology regarding the extraction of the body joint coordinates as well as the experimental setup, analysis, and visualization tool. In Section 3.4 we present the results of our experiments. We discuss them in Section 3.5 and conclude and motivate future research in Section 3.6.

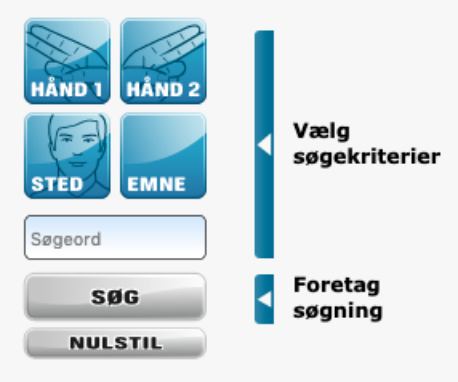

Figure 3.1: Traditional search functionality as seen in the online Danish Sign Language dictionary [30].

3.2 Related Work

Dynamic Time Warping (DTW) is a dynamic programming based time series comparison algorithm to produce a distance metric between two inputs. It has been widely used in the speech recognition domain since the early 1970's [1, 11, 121]. While the original algorithm can be computationally expensive, different variations have been developed over the years to reduce the overall complexity, with most notably the works of Itakura-Parallelogram [82], Ratanamahatana-Koegh-Band [136], and Sakoe-Chiba-Band [139].

As a technique, it has been long-established in the gesture and sign language recognition domain as well [7, 83, 84]. Due to the fact that it is a distance metric, it requires no training and it is a perfect choice for applications where limited training samples are available.

Ten Holt and her colleagues presented an algorithm for Dynamic

Time Warping (DTW) on multi-dimensional time series (MDDTW) to perform classification on 121 gestures recorded with two cameras in stereo position [158]. In Jangyodsuk et al. [84], the authors investigated the use of DTW and Histogram of Oriented Gradient (HOG) to compare a query sign with those in a database of ASL signs using Kinect data. Their results showed an accuracy of 82% in a Top-10 ranking level.

Recent developments in the field of machine and deep learning have lead to advances in sign language and gesture recognition. However, these approaches pose restrictions to their overall applicability as they require large amount of data and computational power in order to be trained. Furthermore, proposed methods for sign language classification have been based on special sensor hardware, such as Microsoft's Kinect presenting additional challenges in their duplicability as well as difficulty in their technical set-up. Our proposed method does not require the use of depth data to extract the pose key-points as this is being held by the pre-trained pose estimation framework OpenPose. This makes our approach suitable for any kind of sign language lexicon.

Most recently, Schneider et al. [144] used Dynamic Time Warping in conjunction with the One-Nearest-Neighbor algorithm and OpenPose to perform classification on six gestures. Their results suggested an accuracy of 77.4%. A major advantage of their methodology is the necessity for very little training data. However, a considerable drawback of their study is that they have only tested a small amount of gestures. As a result, such a pipeline shows a major deterioration of the overall accuracy when an additional gesture is added into the classification task.

Our study re-purposes the work of Schneider et al. by:

- considering signs instead of gestures as inputs in DTW
- extending significantly the number of signs used in the experiment
- adding the finger joints extracted by OpenPose as additional data
- testing whether signing proficiency influences the accuracy of the method

3.3 Methodology

In this section we describe the pose estimation framework (i.e. OpenPose) as well as the apparatus and materials used in this study.

3.3.1 Pose Estimation

OpenPose is a real-time, open-source library for multi-person 2D pose estimation [27]. It can detect body, foot, hand, and facial keypoints. It is a bottom-up approach, meaning that it does not recognize first where a person is in an image and then extract the body joints, but from the detection of the various keypoints, it predicts the overall pose. In general, OpenPose exceeds in performance compared to similar 2D body pose estimation libraries like Mask R-CNN [76] and Alpha-Pose [105]. Its major advantage lies in its high accuracy regardless of the number of people in an image or video.

OpenPose is able to run on different operating systems and hardware architectures while providing all the necessary tools for acquisition, visualization and output file generation. Its output consists of multiple json formatted files containing the pixel x, y coordinates of the body, hand and face joints. In this study only the body and hand predictions were used as the face joints were irrelevant for our purposes.

3.3.2 Pre-processing

The output of OpenPose consists of x,y pixel coordinates. As the people in each frame can potentially be in different locations, it is important to normalize their keypoints. Rotational invariance is omitted in this study as most people are expected to be in an upright position in front of the web camera.

The normalization is done in two steps. Firstly, all the keypoints are translated in such a way that the neck keypoint shifts to the origin at (0,0). To accomplish the shift, the neck keypoint's coordinates are subtracted from all other keypoints. Secondly, the keypoints are scaled in such a way that the distance between the left and the right shoulder keypoint becomes 1. This is achieved by dividing all keypoints' coordinates by the distance between the left and right shoulder keypoint. The scale normalization method is based on previous studies by Celebi et al. [29], Schneider et al. [144], and Ostling et al. [127].

One additional step added to the pipeline is the horizontal flip of the videos when a participant was left-handed. This step is achieved by measuring the average velocity of each hand. In cases where the left hand's velocity is greater than the respective of the right hand, a horizontal flip is applied. Such a process allows an independent

Identifier	**Sign Gloss**
s1	ABOUT
s2	BED
s3	BOOK
s4	CAPTAIN
s5	DREAM
s6	EAT
s7	ELEPHANT
s8	HISTORY
s9	HOTEL
s10	IF
s11	LAPTOP
s12	LATER
s13	LUNCH
s14	MEET
s15	MIND
s16	NEAR
s17	NOSE
s18	OPEN
s19	TALK
s20	TRUE

Table 3.1: List of signs shown to the participants of our experiment

handedness feature of the overall methodology.

3.3.3 Participants

Ten people were asked to participate in the research. Four of them have no experience with sign language whatsoever while the rest are experienced signers. Additionally, they were all informed about the general purpose of the research and gave their consent to participate. This study was approved by the Faculty ethics committee.

3.3.4 Data

Each participant viewed only once a selection of 20 signs from the newly compiled Ghanaian Sign Language lexicon [73]. While the overall lexicon

has more than 1300 signs, we randomly selected 100 of them to be used in our experiments due to time limitations. The order was randomized for each participant to avoid potential biases. A full list can be seen in Table 1. Each video had a resolution of 1000 by 580 pixels at 30 frames per second and lasted approximately 5 (± 2) seconds. Recordings were made with a MacBook Pro's webcam at a resolution of 1280 by 720 pixels and 30 frames per second.

We employ the soft DTW method by Cuturi and Blondel [43] deployed by the tslearn Python package [157] to perform DTW on the normalized trajectories of the dominant hand. Their work takes advantage of a smoothed formulation of DTW that computes the soft-minimum of all alignment costs. In a pilot test, we observed that soft DTW performed better compared to other DTW variants, and was thus used in the rest of the experiment. Furthermore, a DTW variant created by Sakoe and Chiba [139], used by the same Python module, was utilized to measure the distance of the trajectories of all finger coordinates.

Most signs in our lexicon are one-handed where the left hand is inert either by being "absent" or passively fixed at a location. In the two-handed signs, the left hand mostly copies the movement of the right hand. As a result, we employed DTW only on the dominant hand features as the left hand would either be less informative or equally informative.

Finally, the limited resolution of the output from OpenPose had an undesired effect, producing sudden spikes in the signal. This attribute has been previously acknowledged by Schneider et al. [144] and was also present here. The videos in the lexicon were blurry when the hand was moving fast, causing OpenPose to mis-predict the proper joint locations between consecutive frames and resulting in a lack of smoothness in the hand path. To compensate for this behavior, we included two additional steps. Firstly, all the dominant hand's wrist x,y coordinates that had a confidence level lower than 0.3 were deleted. Additionally, we used a median filter with a radius of $r = 3$ for smoothing the remaining signal.

Moreover, we observed a discrepancy in the body joint coordinates predicted by OpenPose due to the high-quality lighting conditions of the GSL lexicon videos. This contrasted with the lower quality, poorly lit videos captured with participants, creating difficulties for the DTW algorithm to function correctly. To remedy this situation, we opted to enhance the lexicon by incorporating data from a randomly selected participant each time we conducted a test. This technique effectively

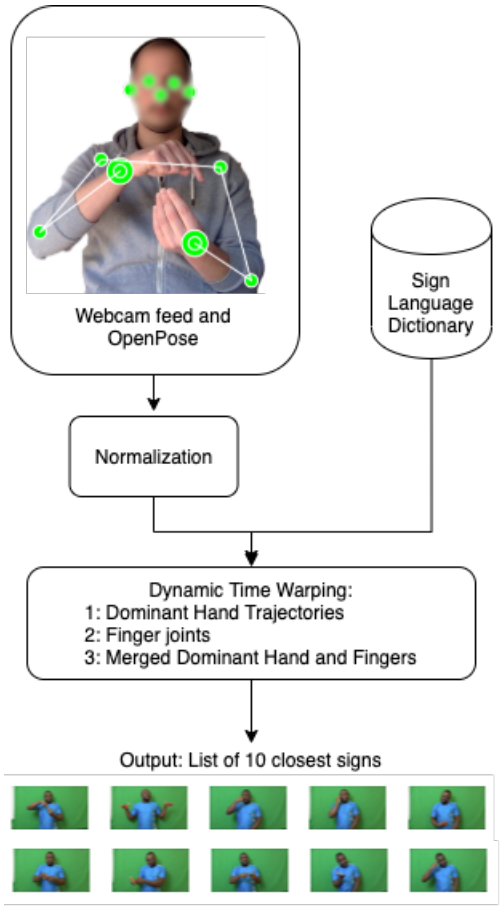

Figure 3.2: Overview of the overall pipeline of our methodology.

introduced a level of noise into the lexicon that mirrored the noise in the participant data, thereby facilitating the DTW algorithm in generating more accurate comparisons. In essence, this strategy created a more equitable data environment for the algorithm, significantly improving its operational efficiency and reliability. As a result, the data of each participant's sign was compared with 120 signs in our database (100 from the GSL lexicon and 20 from another random participant). The overall pipeline can be seen in figure 3.2.

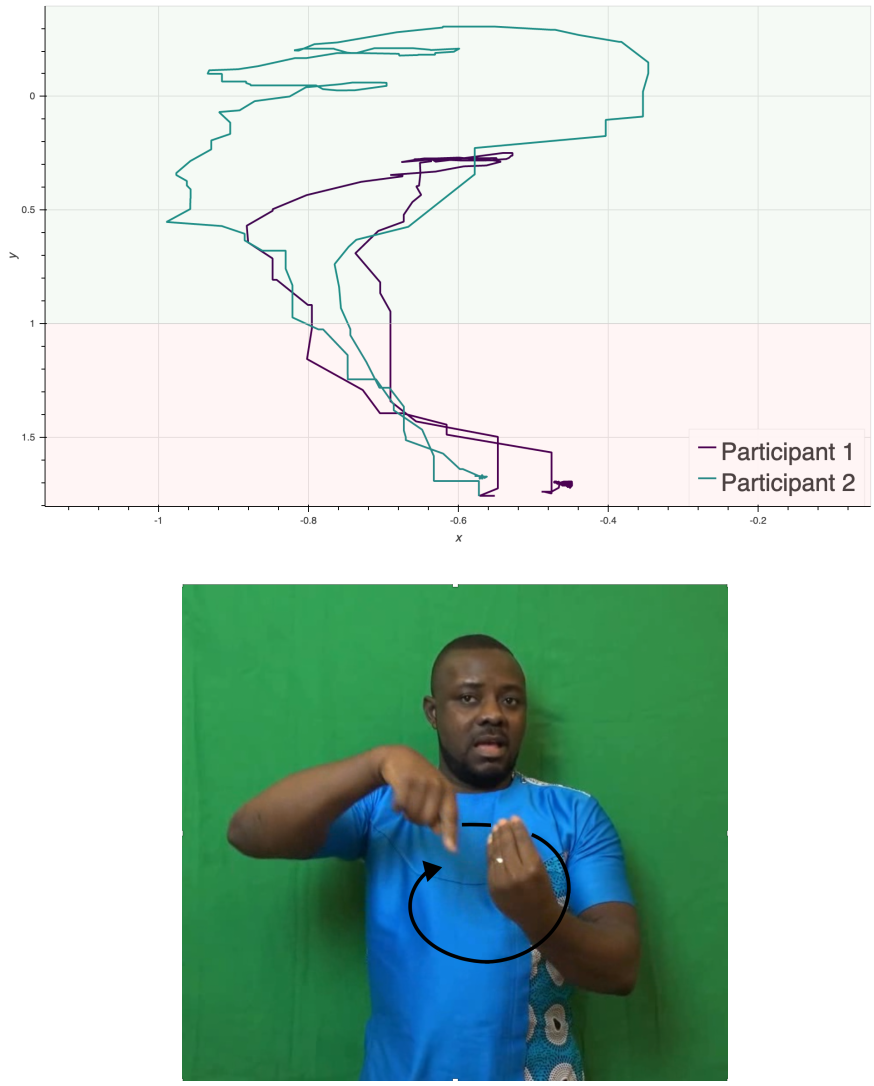

Figure 3.3: Visualization of the dominant hand trajectories between two participants (a) for the sign ABOUT (b).

3.3.5 Visualization

To further explore the outputs of OpenPose and how they are rendered in our methodology, we have created an interactive visualization tool. Developed with the Python module "Bokeh" [17], the user is able to view the extracted dominant hand trajectories from the participants as a whole or individually. As all participants started and ended each sign in the same position, we have color-coded the preparation and retraction phase as red, and the stroke of each sign as green.

While the motivation behind the creation of this tool was solely to verify the output of Openpose and the normalization part in our method, its potential reaches beyond the scope of this study. Such a tool, in combination with the DTW output, can potentially be used as a metric to quantify the variation in the movement and location of signers and sign languages in general. An example of the trajectories of two participants for the same sign can be seen in figure 3.3a. It is evident that one participant produced the sign in a larger space with more distinctive movements. Moreover, it can be deduced that the location parameter is different as Participant 2 made the sign at a higher plane (almost in front of the face) while Participant 1 in front of the torso.

3.4 Results

Table 3.2 presents the overall accuracy of our methodology. Top-k refers to the number of signs a user must look up before finding a correct match. Accuracy indicates whether the target sign is present in the Top-k retrieved signs and is averaged across all participants and signs.

It is evident that the highest accuracy is apparent at a Top-10 rank level at 87%. Furthermore, Top-5 rank shows an adequate accuracy at 74%. Contrary to expectations, using DTW in the joints of fingers extracted by OpenPose did not yield significant results with a highest accuracy at the Top-10 rank at approximately 52%. Merged DTW distances from the dominant hand trajectories and the finger joints also did not generate compelling results.

If only the experienced signers' data is considered then the accuracy at the Top-10 rank raises at 90% and the Top-5 at 78% (Table 3.2 row 4). On the other hand, the accuracy on the non-experienced signers drops at 82% and 67% at the Top-10 and 5 rank respectively (Table 3.2 row 5). Moreover, DTW on the finger's trajectories shows a significant drop at

the Top-10 rank between the experienced and non-experienced signers of approximately 22% (Table 3.2 column 7).

The most striking observation to emerge from the analysis was that four out of 20 signs were consistently recognized with almost 100% accuracy at the Top-1 level rank. These signs were: CAPTAIN, DREAM, ELEPHANT, and OPEN. Such behavior can be justified because these signs exhibit large, distinctive movements and locations that are hard for the DTW to misinterpret.

Condition	Dominant hand trajectory			Fingers' trajectories			Merged trajectories		
Top-k	Top 1	Top 5	Top 10	Top 1	Top 5	Top 10	Top 1	Top 5	Top 10
Accuracy of all participants	0,27	0,74	0,87	0,23	0,40	0,52	0,29	0,55	0,71
Accuracy of exprerienced signers	0,29	0,78	0,90	0,30	0,47	0,61	0,39	0,64	0,79
Accuracy of non-experienced signers	0,25	0,67	0,82	0,12	0,30	0,39	0,14	0,42	0,59

Table 3.2: Sign retrieval accuracy. Top k refers to number of best matches.

3.5 Discussion

In this study we have investigated the use of OpenPose and Dynamic Time Warping as a ranking pipeline to retrieve matching signs from a sign language dictionary. Our results demonstrated that such a task can be achieved with an adequate accuracy rate.

This is in good agreement with the results obtained by Jangyodsuk et al. [84]. Although the accuracy rate does not match the one from Schneider et al. [144], we have tested a larger vocabulary and lexicon. Additionally, we are not aiming to classify each sign but rather create a suggestion ranking system. As such, our results suggest that approximately 9 out of 10 times the matched sign will be present in the first 10 retrieved signs.

Moreover, the results have further strengthened our hypothesis that signing proficiency is an influencing factor for classification efforts. Although our sample size was limited there was a significant drop in the accuracy rates between the experienced and non-experienced signers. The former, produced well structured signs matching more appropriately the ones from the lexicon, which made DTW perform in a more excellent matter.

Our research failed to account for the low values of accuracy on the finger joints. This was probably as a result of the low performance of

OpenPose in accurately predicting the finger joints due to low lighting conditions in the videos. It was often the case that joint predictions would disappear between frames or mis-predicted in wrong locations. Thus, caution must be exercised when OpenPose is being used for such trivial tasks.

3.6 Conclusion

To sum up, we have obtained satisfactory results demonstrating the use of OpenPose and Dynamic Time Warping for a new, sign-based search functionality in reduced sign language dictionaries. We showed that our "find the sign" methodology can be used as a suggestion tool for sign retrieval in a small lexicon by using only the trajectory of the dominant hand. Additionally, our research has highlighted the importance of considering the level of signing proficiency when it comes to classification tasks. The significance of this study lies on the fact that the methodology in question can be easily used in any kind of sign language lexicon, irrespective of its quality and language. Additionally, no prior training of any kind of model is required. As such, this approach, in combination with the developed visualization module, has the potential to be used also as a metric tool to quantify the variation between signers and overall languages.

Furthermore, a number of things is left for future work; first and foremost, to investigate how extracted finger joints can be utilized more efficiently in the overall pipeline. Moreover, different variants of the original DTW algorithms need to be tested. Finally, we intend to evaluate the use of other pose estimation frameworks, such as PoseNet, to further enhance the web and mobile user-friendliness of the method used.