2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW) | 979-8-3503-0261-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSPW59220.2023.10193012

PROCRUSTES-DTW: DYNAMIC TIME WARPING VARIANT FOR THE RECOGNITION OF SIGN LANGUAGE UTTERANCES

Nikolaos Arvanitis, Evangelos Sartinas , Dimitrios Kosmopoulos

University of Patras
Computer Engineering & Informatics Department
Rion-Patras 26504, Greece

ABSTRACT

In this paper we present a method for classifying sign language videos using a variation of dynamic time warping with the Procrustes distance, which is a shape similarity measure appropriate for handshape recognition. We initially extract features from the dominant hand. We then classify the signs using a nearest-neighbor scheme which includes a dynamic time warping variant, which we dub Procrustes-DTW and which is suitable for extracting a similarity measure of sequences of handshapes only. To reduce computational costs it is combined with a curvature-based summarization scheme with a reasonable sacrifice in accuracy. We verify experimentally our approach on a custom dataset and we show its merit compared to the conventional dynamic time warping approach.

Index Terms— Sign Language Recognition, Sign Language Video Summarization, Procrustes-DTW

1. INTRODUCTION

Sign Languages (SLs) are the main communication means among the members of the Deaf community. The development of tools and algorithms that would automatically recognise or translate SLs to spoken languages and vice versa would be of great utility for the communication between deaf and hearing individuals, e.g., for the provision of medical services [1] or access to cultural sites and related content [2] or for SL teaching [3].

In the literature the SL translation problem is typically treated as a sequence-to-sequence problem. The input sequence is the visual depiction of the signer and the output sequence is the translated text or gloss sequence. Deep encoder-decoder architectures are commonly employed for this purpose (e.g., [4], [5], [6]).

However, the above approaches suffer from two major problems:

- The amount of training data is typically too small to capture in a statistically sufficient way the different aspects of the language (vocabulary, syntax, morphological phenomena etc.). Indeed, SLs are minority languages with lack of annotated datasets.
- The videos of the signers may include a lot of video frames, which require rather long processing if we were to consider them all.

In this paper we contribute towards SL recognition considering these issues. More specifically, we seek to recognize different SL video sequences, which is a common problem in a SL teaching context [3], or in a scenario of content access, like cultural content, e.g., [2].

A large number of SL videos would enable the use of some popular neural models on sequence processing. But due to the lack of a sufficient amount of annotated data for learning purposes, we make use of techniques that do not include training as a process of learning input representations and parameter tuning.

In this study, we capitalize on our previous work on SL summarization to extract only the most relevant and meaningful video frames [7]. From the dominant hand in those frames, we extract the hand features; then we classify the sequences using a nearest-neighbor scheme, which includes a dynamic time warping variant, dubbed Procrustes-DTW. This DTW variant turns out to be suitable for extracting a similarity measure for handshape sequences. We are able to classify SL sequences solely based on this similarity of handshapes.

2. RELATED WORK

SL Summarization The summarization of SL videos in the recognition context is part of our investigation, so we give the related context. In [8] the region of hands and face are segmented using skin color and then modeled using Zernike moments. The second derivative of the moment norm may be used to extract the keyframes, assuming these are the turning points in the overall motion. In [9], the frames corresponding to the Maximum Curvature Points (MCPs) of the global tra-

The first author was funded by T2EDK-00982 SignGuide project, RESEARCH-CREATE-INNOVATE, by the Greek Secretariat for Research and Innovation and the EU.

jectory are proposed to be taken as the keyframes for the compression of the sign's video and for solving the corresponding sign classification problem. Authors of [10], [11] used centroid and solidity distance thresholding between consecutive frames to decide if a frame is a keyframe or not. However, the keyframes extracted in this way are not necessarily the most important ones from the point of view of the semantics and can be just the result of transient motion before reaching the next semantically important hand pose.

In our previous work [7] we presented a keyframe extraction scheme based on the wrist motion using differential geometry. More specifically, the time (t)-parameterized Frennet-Serret frame for tracking the signer's wrist is used and the curvature of the trajectory, is proposed for the identification of the SL video keyframes. Specifically, a video frame is characterized as keyframe if on that time instance the t-parameterized curvature function attains a maximum value. That scheme compared favourably to the previous ones and we employ it for our keyframe extraction. We utilised this summarization technique as the more suitable to our methodology, since it has been demonstrated in [7] that high norm values of the curvature's derivative signify in a satisfactory way the most important semantics of the utterance.

SL Recognition First approaches utilised gloves or sensors on signers to achieve recognition. More recent studies focus on the recognition on a continuous stream of signing using vision-based techniques. The advances in Machine Learning introduced CNN, RNN and other sequence neural models with the limitation of the little availability of data [12], [13].

SL Recognition using DTW Dynamic Time Warping (DTW) has been widely used in SL recognition due to its ability to measure similarity between sequences with different lengths and temporal variations. Several studies have been conducted to explore the potential of DTW in SL recognition mainly by exploiting the trajectory of the hands on its own, or in combination with some simplistic representation of the hand shapes.

For example, the authors of [14] reconstructed the trajectory using Simpson adaptive algorithm to perform trajectory recognition using a generalized linear regression based optimization of DTW. In [15], DTW was used for sign trajectory similarity in combination with a handshape representation using Histogram of Oriented Gradients was employed; however, this is a rather crude representation of handshapes. Tang et al. [16] proposed a way to segment continuous trajectory by combining templates and velocity information to spot the beginning and ending points in hand gesture trajectories. The study in [11] suggested the use of Zernike Moments as shape descriptors to extract feature vectors and recognize gestures. It may be an interesting approach, but more representative features would be desirable. In [17], the DTW distances are used as a feature vector by combining a trajectory representation with the hand orientation. Then the classification was performed using LDA, SVM and KNN.

Our work is the first that introduces a way to match SL handshape sequences using Procrustes Analysis as a shape distance into DTW and perform nearest neighbour search to recognise SL utterances on summarized videos.

3. METHODOLOGY

The SLs are notorious for their scarce datasets, so a methodology based on a nearest neighbor scheme, which does not require data-intensive training appears to be an attractive choice. The overall approach is shown in Fig. 1 and includes a feature extraction and a classification step. Firstly, a struct is created, which contains every label and each label contains its corresponding videos. The process involves summarizing every video by extracting its keyframes and applying the Mediapipe Holistic module [18] to each keyframe, resulting in the extraction of 21 3-D landmarks that reveal shape information. This results in a representation of the video in the form of a series of hand-landmarks.

The same procedure is applied to every test video to extract the handshape keyframe series. A test video is compared to all others in the labels struct in terms of similarity. Procrustes-DTW is used as a distance between a test video and a reference one. The distances of a test video are stored temporarily in a list, and the minimum distance from that list is used to assign a class to the test video. This class corresponds to the nearest neighbor's video label.

3.1. Summarization and Feature Extraction

In a SL video only a small fraction of the frames is actually needed to understand the meaning. In this work, we investigate a keyframe extraction scheme based on the wrist motion using differential geometry. It identifies keyframes depending on whether or not the time parameterized curvature function of a signer's wrist attains its maxima values (see [7] for details).

This keyframe extraction scheme, made it possible to reduce the computational costs and was able to smooth the trajectory and speed variations among different signers that naturally arise during a realistic SL utterance.

We represent every keyframe as a list of 3-D hand landmarks as they are extracted by Mediapipe ([19]), see Fig. 2. We used Mediapipe Holistic Module and kept its 3-D coordinates of 21 the landmarks of the hand, that is the x-value, the y-value and an estimation of z-value regarding the position of the hand in the frame.

3.2. Procrustes-DTW distance

Procrustes analysis, is a statistical tool useful for the analysis of the distribution of shape vectors [20]. It helps to compare two shapes and extract their distance as a measurement of their similarity. To be able to be compared under Procrustes

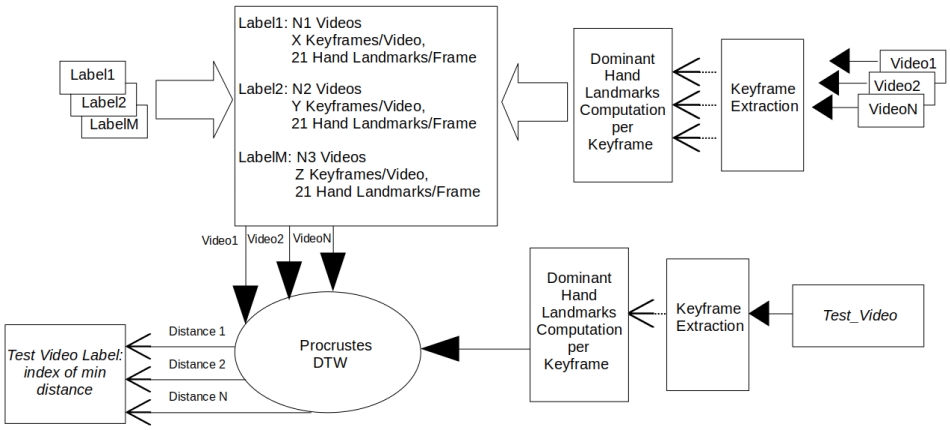
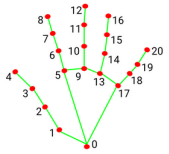
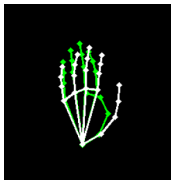


Fig. 1: The figure depicts the overall approach which is described in the first two paragraphs of Section 3.

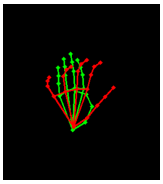


(a) The 21 Mediapipe hand landmarks [18]

(b) Example of hand landmarks extraction.



(c) Two matching handshapes aligned



(d) Two non-matching handshapes aligned

Fig. 2: Above: Mediapipe Hand Landmarks, Below: Superimposition of two handshapes

scheme, the two shapes firstly need to be superimposed. This superimposition includes an optimal translation, rotation and uniform scale of the two objects. An example of two superimposed handshapes is shown in Fig. 2.

In order to find the distance between two shape matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we need to solve the Orthogonal Procrustes Problem, which is related to matrix approximation. Specifically the problem is solved when the orthogonal matrix $\boldsymbol{\Omega}$ is found that maps optimally $\boldsymbol{A}$ to $\boldsymbol{B}$:

$$\boldsymbol{R} = \underset{\boldsymbol{\Omega}}{\operatorname{argmin}} \|\boldsymbol{\Omega A} - \boldsymbol{B}\|_F \;\; s.t. \;\; \boldsymbol{\Omega}^T\boldsymbol{\Omega} = \boldsymbol{I} \tag{1}$$

It is equivalent to solving the closest orthogonal approximation problem, that is finding the nearest orthogonal matrix $\boldsymbol{R}$ to a given matrix $\boldsymbol{M}$, so that $\boldsymbol{M} = \boldsymbol{BA}^T$

$$\min_{\boldsymbol{R}} \|\boldsymbol{R} - \boldsymbol{M}\|_F \;\; s.t. \;\; \boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I} \tag{2}$$

Then Singular Value Decomposition is applied on $\boldsymbol{M}$ $(\boldsymbol{BA}^T)$, $\boldsymbol{U\Sigma V}^T = SVD(\boldsymbol{BA}^T)$ to find matrix $\boldsymbol{R}$,

$$\boldsymbol{R} = \boldsymbol{UV}^T \tag{3}$$

Dynamic Time Warping optimally aligns two time-series that may differ in terms of phase and calculates their distance. Traditionally, DTW works with the Euclidean distance. But the Euclidean distance behaves poorly in the case of handshape time-series.

This is because the handshape representation includes 63 elements, i.e., a list of 21 3-D hand landmark coordinates. In such cases the Euclidean distance suffers from the curse of dimensionality, see e.g., [21], [22].

Replacing Euclidean distance with Procrustes distance is a very reasonable choice for handshape time-series, as Procrustes analysis is destined to be used for shape alignment.

So, having computed matrix $\boldsymbol{R}$ of eq. 3, which is optimal in minimizing eq. 2, we extract the distance of the two handshape matrices, using eq. 1 with an inserted normalization scale factor:

$$s = tr(\boldsymbol{\Sigma})/tr(\boldsymbol{B}^T\boldsymbol{B}) \tag{4}$$

The Procrustes-DTW distance is then formed by the following equation:

$$d = \|\boldsymbol{RBs} - \boldsymbol{A}\|_F \tag{5}$$

A nearest neighbour search using Procrustes-DTW distance (see Fig.3) can then be performed to classify a summarized test video (see Fig. 1).

4. EXPERIMENTS

The video dataset used was created by us for the purposes of access to cultural and educational content. It was created

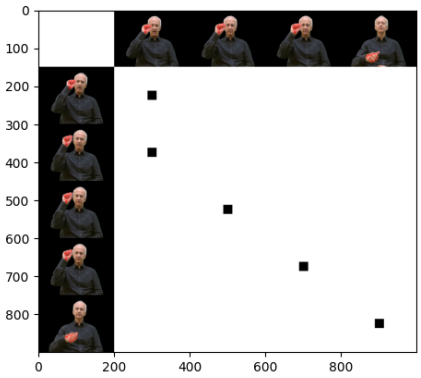


Fig. 3: Example of Procrustes-DTW alignment path of two same class hand feature-series.

using Microsoft Kinect sensors with video recording rate of 30 fps and frame dimensions 1920x1080. It consists of short time (3-6 seconds) videos in which simple Greek glosses or phrases are signed such as visual queries for cultural content in a museum, the letters of the alphabet, the numbers, the members of a family, some simple phrases combining the aforementioned glosses, etc. The number of signers varies from one to nine among labels.

We used 559 videos which correspond to a mix of 145 class labels, mainly glosses and some small phrases. Every gloss label does not have a standard number of video instances but we excluded gloss labels with only one video instance as we would not be able to test them. The amount of samples per class ranges between 2 to 9 videos.

In Table 1 there are shown some label examples of the dataset used along with the amount of videos of each label. We chose to use this dataset because of the high quality conditions that the videos were recorded on, that is the relatively high frame rate and the very few blurry frames. A sufficient high frame rate reduces blur and is important for the Mediapipe to return good hand landmarks. The RWTH Phoenix Weather dataset [23] has been widely used in SL studies, but we could not use it as its frames are blurry and Mediapipe fails to compute precise features.

In order to compare our proposed method in terms of classification accuracy, we carried out six different similarity search experiments. The first one utilizes classic DTW with euclidean distance on hand trajectory feature vectors, that is the dominant hand wrist landmark. Also we employed classic DTW (frobenius norm was used since the data are matrices, proportionally to euclidean distance) comparing sequences of all dominant hand landmarks.The third case is our proposed methodology, a Procrustes-DTW similarity search on hand landmarks. In both scenarios we also included the non summarised video case, so features were extracted from every frame during a signing. Classification accuracy results and mean classification time are shown in Table 2. Note that

Label	# Videos
Α-Ω (Greek alphabet)	3
Χρυσομέταξο Ύφασμα (Goldsilk fabric)	9
Πόσα χρυσά στεφάνια υπάρχουν· (How many gold wreaths are there?)	9
Κρανίο Πετραλώνων (skull of Petralona)	9
Πώς διατηρήθηκε η μούμια· (How was the mummy preserved?)	9
Οστρέινα Βραχιόλια (Oyster Bracelets)	9
Άνοιξη (Spring)	5
Και τέταρτο (Quarter past)	3

Table 1: Examples of dataset labels. In parenthesis, there is the English translation of each label.

experiments were executed on a Intel(R) Core(TM) i7-8700K CPU @ 3.70GHz CPU.

Method	Accuracy	Mean Time
DTW on hand traj. w/o video summ.	64.37%	7865ms
DTW on hand traj. with video summ.	53.51%	745 ms
DTW on all landmarks w/o video summ.	95.54%	40114ms
DTW on all landmarks with video summ.	80.98%	804 ms
Procrustes-DTW w/o video summ.	97.02%	53132ms
Procrustes-DTW with video summ.	86.38%	5712 ms

Table 2: Classification (top-1) accuracy results and mean classification time.

Considering the results of Table 2, classification using Procrustes-DTW as a distance performs clearly better than the other two more conventional approaches, for the recognition of a gloss or a phrase. The higher accuracy is noticeable when not using summarization. However regarding the mean classification time, in a real-time scenario summarization seems essential. A drawback of the suggested method is that since it compares shapes it can not take into acount other useful features (such as the dynamics of the motion of the hand).

5. CONCLUSION AND FUTURE WORK

The idea of using Procrustes-DTW to classify SL videos seems very promising without requiring data-intensive training, which in most cases are not available anyway for SL. The summarization offers a reasonable trade-off between high performance and efficiency, since the Procrustes optimization has rather high computational costs. We have seen that it is possible to obtain decent results by employing only the shape of the dominant hand. We did not investigate efficient implementations of k-NN, like the k-D tree which are expected to reduce the computational cost from linear to logarithmic, but we plan to do so in the near future.

6. REFERENCES

[1] E.V. Pikoulis, A. Bifis, M. Trigka, C. Constantinopoulos, and D. Kosmopoulos, "Context-aware automatic sign language video transcription in psychiatric interviews," *Sensors*, vol. 22, no. 7, 2022.

[2] Kosmopoulos D. et al, "Museum guidance in sign language: The SignGuide project," 2022, PETRA '22, p. 646–652.

[3] Y. Zhang, Y. Min, and X. Chen, "Teaching chinese sign language with a smartphone," *Virtual Reality & Intelligent Hardware*, vol. 3, no. 3, pp. 248–260, 2021.

[4] A. Voskou, K. P. Panousis, D. Kosmopoulos, D. N. Metaxas, and S. Chatzis, "Stochastic transformer networks with linear competing units: Application to end-to-end SL translation," in *ICCV*, 2021, pp. 11926–11935.

[5] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Multi-channel transformers for multi-articulatory sign language translation," in *Computer Vision – ECCV 2020 Workshops*, 2020, pp. 301–319.

[6] N. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *CVPR*, 03 2020.

[7] E. Sartinas, E. Z. Psarakis, K. Antzakas, and D. I. Kosmopoulos, "A 2-d wrist motion based sign language video summarization," in *BMVC*, 2021, p. 227.

[8] D. I. Kosmopoulos, A. Doulamis, and N. Doulamis, "Gesture-based video summarization," in *IEEE Int. Conf. on Image Proc.*, Sep. 2005, vol. 3, pp. III–1220.

[9] M. Geetha and P.V. Aswathi, "Dynamic gesture recognition of indian sign language considering local motion of hand using spatial location of key maximum curvature points," in *RAICS*. IEEE, 2013, pp. 86–91.

[10] B. Pathak, A. S. Jalal, S. C. Agrawal, and C. Bhatnagar, "A framework for dynamic hand gesture recognition using key frames extraction," in *NCVPRIPG*, 2015.

[11] S. Mathur and P. Sharma, "Sign language gesture recognition using zernike moments and dtw," in *SPIN*, 2018, pp. 586–591.

[12] Dr. M. Madhiarasan and Prof. Partha Pratim Roy, "A comprehensive review of sign language recognition: Different types, modalities, and datasets," 2022.

[13] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef, C. Vogler, and M. Ringel Morris, "Sign language recognition, generation, and translation: An interdisciplinary perspective," in *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, New York, NY, USA, 2019, ASSETS '19, p. 16–31, Association for Computing Machinery.

[14] W. Li, Z. Luo, and X. Xi, "Movement trajectory recognition of sign language based on optimized dynamic time warping," *Electronics*, vol. 9, no. 9, 2020.

[15] P. Jangyodsuk, C. Conly, and V. Athitsos, "Sign language recognition using dynamic time warping and hand shape distance based on histogram of oriented gradient features," in *PETRAE*, 2014.

[16] Jingren Tang, Hong Cheng, Yang Zhao, and Hongliang Guo, "Structured dynamic time warping for continuous hand trajectory gesture recognition," *Pattern Recognition*, vol. 80, pp. 21–31, 2018.

[17] W. Liu, J. Cheng, F. Wei, Y. Liu, C. Li, Q. Chen, and X. Chen, "Chinese sign language recognition based on dtw-distance-mapping features," *Mathematical Problems in Engineering*, June 2020.

[18] "Mediapipe hands," `https://google.github.io/mediapipe/solutions/hands.html`, Accessed: 2023-03-01.

[19] C. Lugaresi et al, "Mediapipe: A framework for building perception pipelines," *CoRR*, vol. abs/1906.08172, 2019.

[20] J.C. Gower and G.B. Dijksterhuis, "Procrustes problems. new york: Oxford university press," *Psychometrika*, vol. 70, 12 2005.

[21] C. Robert Taylor, *Dynamic Programming and the Curses of Dimensionality*, chapter 1, CRC Press, 1993.

[22] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?," in *Database Theory — ICDT'99*, Catriel Beeri and Peter Buneman, Eds. 1999, pp. 217–235, Springer Berlin Heidelberg.

[23] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Comp. Vis. and Im. Underst.*, vol. 141, pp. 108–125, Dec. 2015.