Towards a visual Sign Language dataset for home care services

D. Kosmopoulos[1], I. Oikonomidis[2], C. Constantinopoulos[1], N. Arvanitis[1], K. Antzakas[1], A. Bifis[1], G. Lydakis[2], A. Roussos[2], A. Argyros[2]

[1] University of Patras, Patras, Greece

[2] Foundation for Research and Technology – Hellas (FORTH), Heraklion, Greece

***Abstract*— We present our work towards creating a dataset, which is intended to be used for the implementation of a home care services system for the deaf. The dataset includes recorded realistic scenarios of interactions between deaf patients and mental health experts in their native sign language. The scenarios allow for contextualized representations, in contrast to typical datasets presenting isolated signs or sentences. It includes continuous videos in RGB and depth, which are challenging to analyze and closely resemble real-life scenarios. The research on representation of signs is supported by providing the hand shapes and trajectories for every video using hand and skeleton models, as well as facial features. Furthermore, the dataset may be used for studying the emotional context, since such conversations are typically emotionally charged.**

I. INTRODUCTION

Sign Languages (SLs) are typically the native languages used by the Deaf. Therefore, the access to SLs is essential for the fulfillment of the basic human right of communication for the Deaf. However, there is a universal shortage of interpreters, which very often undermines the right of the deaf to communicate in their native language and may endanger their lives, especially in cases of emergency or serious health incidents. Access to texting may be a solution, but is far from optimal, since most Deaf have problems in reading and writing due to their poor language experiences [12].

We aim to develop an application for the automated interpretation of the Greek Sign Language (GSL) over internet with focus on the health services, which is a common reason to seek for an interpreter. We aim to develop a dataset to represent realistic visual scenarios. Such scenarios are meant to be used for training a machine learning system that will be able to interpret SL phrases, considering the context. Apart from the linguistic point of view, the dataset may be used for the study of emotion in SL, since the content of the conversation between patient and doctor is typically emotionally charged.

The progress in Natural Language Processing via deep learning is significant in the recent years. Therefore, the modeling of SL cannot be complete just by regarding it as a series of visual classification tasks. Recently, the application of deep learning methods to NLP was made possible by representing words as vectors in a low-dimensional continuous space. In that fashion, the word embeddings are static, i.e., each word has a single vector, regardless of context. However, the problem is that all the different meanings of polysemous words have to share the same representation. More recent works, namely deep neural language models such as ELMo [16] and BERT [6] have successfully created contextualized word representations, word vectors that are sensitive to the context in which they appear. Replacing static embeddings with contextualized representations has yielded significant improvements.

The added value of the dataset can be described as:

- Realistic visual dataset on dialogues between patients and mental health doctors using SL within the context of home care service scenarios.
- Annotations including structural information, i.e., skeletal and hand poses, and facial features.
- Annotations including semantic information, i.e., glosses and emotional context.

The rest of the paper is structured as follows. In the next section we present the currently available datasets for sign language interpretations. In section III we present the dataset and its features. In section IV we present the processing we did to extract pose (hands/faces). Section V presents the next steps and section VI concludes the paper.

II. RELATED WORKS

In the past there have been presented several video datasets aiming to support SL analysis and interpretation. Some of them focus on SL Recognition (SLR), while others focus on linguistic analysis. The former ones are typically used by machine learning methods and involve isolated signs or simple phrases, which are repeated many times. However, this setting is not entirely realistic, in the sense that depicting only isolated words or phrases, which eliminates context. The interpretation is a much more complex task involving analysis of continuous signing, which may be challenging due to co-articulation effects, as well as due to context. The latter ones typically include sample signs or natural narrations from which context information may be extracted, however they don't offer enough (or any whatsoever) repetitions from which learning methods can benefit.

The MS-ASL [19] is the largest dataset currently available, and therefore the most appropriate for machine learning. It contains more than 45000 videos from various online sources, which have been processed mostly automatically. It includes a vocabulary of 1000 signs. It used OCR to recognize the subtitles and face recognition to use separate signers in training, testing and validation sets. Four subsets

Cofinanced by the EU and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project: T1EDK-01299 - HealthSign).

including the 100, 200, 500 and 1000 most frequent words were defined. One of the issues is the isolated nature of the presented words or sentences, which may eliminate the context that could carry useful information. Another problem could be the lack of depth information, which may present difficulties to methods based on depth for hand tracking.

The RWTH-PHOENIX-Weather 2014 [10], is probably the most highly cited dataset in the SLR literature. This is due to the fact that it is free and is the most realistic, since it is taken from real weather reports in the German SL. It contains 1980 sentences and 9 signers. However, the data are limited to low resolution RGB videos, with a lot of blur, due to interlaced frames, which do not allow for accurate hand tracking.

The SIGNUM [1] is a very interesting dataset including 780 sentences of general content, performed by 20 signers in German SL. It is close to our work, due to the continuous nature of the signing. The high-quality videos are a significant plus. However, the sentences are isolated, which prohibits the use of tools that make use of context. Moreover, only RGB videos are available, which may impede pose extraction. It is not available for free, which partially explains its rather infrequent use by the SLR community.

The BosphorusSign [2] is a dataset that involves 855 signs and phrases from the domains of health, finance and some common phrases in Turkish SL. RGB and depth videos were acquired. The annotation includes glosses, sign time intervals and HamNoSys notation. The dataset includes very interesting features, and appears to be the closest to the domain of our dataset. Unfortunately, it has not been fully released to the research community and therefore it has been rarely employed in SL modeling, mainly by its authors.

The CSL dataset [9] covers 100 continuous Chinese SL sentences. 50 signers were used, each of them producing the signs five times. The videos include RGB and depth. It has a vocabulary of 178 signs. Again, the sentences are unrelated.

The RWTH-Boston-400 dataset [7] contains 843 sentences of the American SL, divided into 633 for training, 106 for validation and 104 for testing. These are RGB videos that are not connected to each other.

The above datasets have played a significant role in SL analysis and will continue to do so. What seems to be missing is having multiple instances of the same scenarios performed by several signers, so that the conversation or narration context may be captured. Furthermore, the emotional analysis, which seems to offer context information in dialogue systems (e.g., [8]) and may be of utility in interpretation as well, is also missing from all related datasets. Last but not least, structural information such as user hand/body pose and facial expressions are not part of those datasets, despite their usefulness. Table I gives an overview of the datasets and their features compared to ours.

III. The Dataset and its Features

Instead of offering isolated signs or sentences as most SL-datasets do, we provide several instances of scenarios representing realistic interactions between mental health professionals and deaf patients. We record the patients' part in videos. This is expected to be of utility to researchers seeking to combine the structural analysis of SL with semantics and context information. We also provide emotion tags for every sentence. One last contribution is the representation of SL using skeleton and hand models[1].

Fig. 1. The acquisition system setup, including Kinect sensor, camera and monitors.

A. Corpus compilation

The corpus focuses on the health domain, and specifically on the treatment of patients suffering from stress and depression. As the source of stress is often related to life events, and it also affects everyday behaviour, we obtained a balanced vocabulary consisting of medical and quotidian terms.

Let us note that initial attempts of gathering relevant text from online forums gave poor results, because that content was created without the immediate interaction between a patient and a mental health expert. To tackle this issue, we requested the assistance of two experienced mental health experts, who produced scripts of dialogues between an expert and a patient. The scripts are modified versions of real depression and stress medical cases.

The current version of the corpus contains 21 recorded scripts in Greek Language (GL), each of them performed by 8 speakers of the Greek Sign Language (GSL). Six were native speakers and the rest of them were experienced interpreters (no significant differences were noticed between these two categories). Four were male and four female, with an age range between 20 and 55.

We recorded high quality video using a machine vision camera, as well as depth information using the Kinect 2 device. The post-processing of the videos included their manual annotation with glosses of the GSL, as well as the keypoints of the hands and the face of the signer, which were detected by appropriately adapted variations of the works [14] (hands), [4] (faces).

Furthermore, each sentence was annotated for emotion using the following labels: "happiness", "anger", "fear", "disgust", "sadness", "surprise", "neutral". The annotation considers facial expressions, motion, as well as the semantics of the uttered sentences.

[1]Sample data for one scenario and one signer can be downloaded from: https://doi.org/10.7910/DVN/FC734E

TABLE I

COMPARING THE MOST POPULAR CONTINUOUS SL DATASETS AND THEIR FEATURES TO OURS.

Dataset	Modalities	SL	Signers	Vocabulary	Domain	Annotation	Pose	Context
MS-ASL [19]	RGB	American	222	1000 signs	general	gloss	×	×
RWTH-PHOENIX-Weather 2014 [10]	RGB (210×260, 25fps)	German	9	7k sentences 1081 signs	weather	glosses, approx. borders	×	×
BosphorusSign [2]	depth, RGB, human pose, user mask	Turkish	10	855 phrases	health, finance	border, gloss, HamNoSys	×	×
SIGNUM [1]	RGB (780×580, 30fps)	German	20	780 sentences 450 signs	general	gloss	×	×
CSL [9]	RGB (1920×1080, 30fps) depth (512×524)	Chinese	50	100 sentences 178 signs	general	sentence	✓ skeleton	×
RWTH-Boston-400	RGB (648×484, 60fps)	American	4	843 sentences 483 sign classes	general	gloss, separated	×	×
OUR (currently)	RGB (1920×1080, 30fps), depth (512×524), RGB (1232×1028, 55fps)	Greek	8	945 sentences 806 signs	stress depression	gloss, separated	✓ skeleton, hands, face	✓
OUR (target)	RGB (1920×1080, 30fps), depth (512×524), RGB (1232×1028, 55fps)	Greek	≥15	≥4000 sentences ≥1500 signs	stress, depression	gloss, separated	✓ skeleton, hands, face	✓

B. Recording setup

To control the inevitable variability of expressions between different signers, we simplified the scripts by breaking down each sentence into several short sentences and simple phrases of no more than ten words. Then we transcribed the text using the glosses of GSL. The glosses were dictated to a master signer, and we recorded her signs. The resulting videos were utilized as the reference material for the signers during the actual data collection.

For the recording we used first a machine vision camera, namely the MC050MG-SY model by Ximea, which delivers RGB videos of up to 60fps to capture fast motion and reduce blur. Second, the Kinect 2 device by Microsoft, which allows for synchronous RGB video and depth data capturing. The camera was positioned 1.0 m away from the signer, who was seated in front of a green screen. The Kinect sensor was placed just above the camera. Above both of them there was a TV screen, showing the reference video with the master signer. This arrangement is depicted in Figure 1.

The recorded videos were annotated with GSL glosses using the ELAN [11] software[2], which produced XML files consisting of the intervals of action and the corresponding glosses. Moreover, we tracked both hands and the face of the signer, and recorded a set of 3D keypoints for each body part. The keypoints of a set correspond to predefined positions of the relevant body part. Our software produced JSON files, listing the sets of detected keypoints per frame.

C. Features

The dataset so far consists of 21 scripts. It includes 1029 simple sentences, signed by 8 signers. Excluding repetitions, there are 945 unique sentences. Moreover, the GL vocabulary contains 1374 unique words, while the total number of words is 6319. Thus the average length of a sentence is 6.1 words. Finally, the words are forming 3558 unique 2-grams, 3841 unique 3-grams and 3292 unique 4-grams.

Regarding the statistics of the glosses, the numbers are similar. The GSL vocabulary contains 806 unique glosses, while the complete corpus contains 2619 glosses. Thus the average length of a sentence is 3.9 glosses. Finally, the glosses are forming 1666 unique 2-grams, 1337 unique 3-grams and 870 unique 4-grams.

We made 432 shots, resulting in a total of 6 hours of video for each modality. The total size of the content recorded by Ximea was 123GB, and by Kinect was 283GB. The Table II accumulates the properties of the video files. We aimed to record high quality data, with an average bitrate of 25 Mbps, suitable for research purposes. An appropriate reduction of the bitrate, close to 1.5 Mbps, could yield a dataset better resembling the conditions of a video conference.

TABLE II

THE PROPERTIES OF THE VIDEO FILES

Stream	Container	Codec	Resolution	Framerate
video by Ximea	AVI	H.264	1232 × 1028	55 fps
video by Kinect	QuickTime	H.264	1920 × 1080	30 fps
depth by Kinect	QuickTime	JPEG2000	512 × 424	30 fps

IV. POSE EXTRACTION

One of the basic features of the dataset is the availability of 3D hand pose for both hands, and 3D location of facial landmarks in each frame. It can be used by any researcher who wants to by-pass the very challenging problems of hand tracking and facial landmark estimation.

For the annotation of the 3D hand pose, we automatically process each frame of the dataset using an appropriately adapted version for 3D hand pose estimation, as presented in [14]. We follow standard conventions for the keypoints of the body, hands, and face. Namely, for body and hands we follow the conventions adopted by Cao et al. [3] and Simon et al. [18] in the OpenPose line of work, and for faces the convention followed by Deng et al. [4].

This approach tackles the problem in three main steps: hand detection in the input image, 2D keypoint estimation, and estimation of the 3D position of the same keypoints in a last refinement step. More specifically, in that work, the input image is processed to detect all instances of hands, using a

[2] http://tla.mpi.nl/tools/tla-tools/elan

Fig. 2. Visualization of the annotations in the dataset. From left to right, the 2D body pose, 3D pose of the left hand, and 3D face structure are visualized for different frames of the dataset.

retrained YOLO v2 object detector [17]. In the next step, each detected hand instance is processed in turn as follows: firstly, it is cropped as indicated by the object detector, and resized to a standard image size. Then, it is provided the 2D hand keypoint detector by Simon et al. [18]. Finally, in the third step, the 2D keypoints detected in the previous step are used as input in an optimization process that minimizes the projection distance of hypothesized 3D keypoints from the corresponding, observed keypoints.

For this annotation, we change the first step of that method, responsible for the detection of hand instances (if any), and cropping the input image according to each detection. In [14] this step is performed by an appropriately trained hand detector, based on YOLO v2 [17]. Here, this is exchanged by wrist detection based on OpenPose body skeleton estimation [3]. Since we know that the person is well placed within the image frame, we can rely on the detection of OpenPose. In the rare case of body pose detection failure, we resort to tracking the hand position for a few frames, by reusing the last known good position, and enlarging the crop area.

Given that we also capture depth data, we are evaluating 3D hand keypoint estimation methods. This can result in more robust estimation of hand keypoints, given the more informative nature of depth maps compared to color images.

Complementary to hand pose, useful contextual information is conveyed through facial expressions. We use a method based on the approach adopted by Deng et al. [4] for the semi-automatic annotation of the Menpo dataset, presented in that work. More specifically, similarly to the case of hands, the 2D location of facial landmarks is first estimated [5], and then an optimization process is employed to estimate the 3D positions that best match the observed 2D landmarks, as described in Section 2.4 of the Menpo work [4].

Sample annotations are visualized in Figure 2. Rows correspond to different dataset frames, and columns to different annotation types. From left to right, the first column corresponds to the estimated 2D body landmarks, as computed by the work by Cao et al. [3]. The second column to the estimated 3D position of the left hand. And the third column to the estimated 3D posture and deformation of the face.

V. Next Steps

Currently, we are in the process of acquiring 30 more scenarios. The final version of the dataset is planned to include approximately 100 scenarios with more than 15 signers. This is expected to create a rich dataset which will include many typical cases of depression and anxiety. It is also expected to create an adequate corpus for modeling via machine learning methods and to exceed the largest number of classes (signs) currently available in the RWTH-PHOENIX-Weather dataset.

The annotation is one of the challenges to be addressed, since typically 90% of the dataset development time is used by human annotators. To this end, we are currently evaluating semi-automated methods that make use of common subsequences identification among videos of the same scenarios performed by different users (see e.g. [13], [15]). The methods are expected to provide approximate time intervals, which will be verified by human annotators. This way we expect to reduce the annotation time to half. Furthermore, we plan to include some performance benchmarks using baseline methods to facilitate comparison in the future.

VI. Conclusions

We have presented our work in progress concerning the creation of a dataset for the Greek SL. We have highlighted the need for SL datasets that will not only include isolated signs or sentences, but that will incorporate full narrations or dialogues. This offers the possibility to create new methods that capitalize on recent advances on contextualized representations from the field of NLP. Moreover, we include information about the expressed emotions and the facial features, as well as hand and body pose which, to our knowledge are unique. Finally, we have provided information about the setup and the features of the dataset, as well as our next steps towards its completion.

References

[1] U. Agris and K.-F. Kraiss. Towards a video corpus for signer-independent continuous sign language recognition. In *Proceedings of Gesture in Human-Computer Interaction and Simulation. International Gesture Workshop*, Lisbon, 2007.

[2] Necati Cihan Camgöz, Ahmet Alp Kındıroğlu, Serpil Karabüklü, Meltem Kelepir, Ayşe Sumru Özsoy, and Lale Akarun. BosphorusSign: A Turkish sign language recognition corpus in health and finance domains. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1383–1388, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[4] Jiankang Deng, Anastasios Roussos, Grigorios Chrysos, Evangelos Ververas, Irene Kotsia, Jie Shen, and Stefanos Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *International Journal of Computer Vision*, 127(6-7):599–624, 2019.

[5] Jiankang Deng, George Trigeorgis, Yuxiang Zhou, and Stefanos Zafeiriou. Joint multi-view face alignment in the wild. *IEEE Transactions on Image Processing*, 28(7):3636–3648, 2019.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7] Philippe Dreuw, Carol Neidle, Vassilis Athitsos, Stan Sclaroff, and Hermann Ney. Benchmark databases for video-based automatic sign language recognition. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*, 01 2008.

[8] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Ni Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. pages 154–164, 01 2019.

[9] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *AAAI*, 2018.

[10] O. Koller, J. Forster, and H. Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125,.

[11] H. Lauberg and H. Sloetjes. Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods, Instruments, and Computers*, 41(3):841–849, 2009.

[12] Lori A. Pakulski and Joan N. Kaderavek. Facilitating literacy using experience books: A case study of two children with hearing loss. *Communication Disorders Quarterly*, 25(4):179–188, 2004.

[13] Costas Panagiotakis, Konstantinos Papoutsakis, and Antonis A Argyros. A graph-based approach for detecting common actions in motion capture data and videos. *Pattern Recognition*, 79:1–11, July 2018.

[14] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 436–445. IEEE, 2018.

[15] Konstantinos Papoutsakis and Antonis A Argyros. Unsupervised and explainable assessment of video similarity. In *British Machine Vision Conference (BMVC 2019)*, Cardiff, UK, September 2019. BMVA.

[16] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[17] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[18] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017.

[19] Hamid Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. In *The British Machine Vision Conference (BMVC)*, September 2019.