CorSiL: A Novel Dataset for Portuguese Sign Language and Expressiveness Recognition

Pedro M. Ferreira[1]
pmmf@inesctec.pt
Inês V. Rodrigues[1]
ines.vigario@gmail.com
Ana Rio[4]
anaespinheira@hotmail.com
Ricardo Sousa[2]
rsousa@rsousa.org
Eduardo M. Pereira [1,3]
ejmp@inesctec.pt
A. Rebelo[1]
arebelo@inesctec.pt

[1] INESC TEC
Rua Doutor Roberto Frias, 378, 4200-465
Porto, Portugal

[2] INEB - Instituto de Engenharia Biomédica
Rua do Campo Alegre, 823, 4150-180
Porto, Portugal

[3] Faculdade de Engenharia da Universidade do Porto
Rua Dr.Roberto Frias, s/n, 4200-465
Porto, Portugal

[4] Agrupamento de Escolas Eugénio de Andrade
Rua Augusto Lessa, 4200-098
Porto, Portugal

Abstract

One of the main challenges in the development of any automatic recognition system, specially in the sign language field, is the availability of suitable ground-truth data. In this paper, a novel video-based database, called CorSiL, is proposed. It comprises two major components: (i) a Portuguese Sign Language dataset and (ii) a duo-interaction dataset, between Deaf and/or hearing people. The database can be used for different purposes like Sign Language recognition tasks or emotion/expressiveness recognition from body language. The whole database along with manual annotations, including signs and body parts, will be available for research and benchmark purposes.

1 Introduction

Sign Language (SL) is the medium of communication between hearing impaired people. It is a multimodal language, since it involves manual (i.e. hand gestures) and non-manual signs (i.e. facial expressions, head motion, pose and body movements). Therefore, emotions and expressions play a crucial role to convey meaning. As SL is commonly used just within deaf communities, with its own lexicon and grammar, the majority of hearing people are unfamiliar with the SL. In order to overcome the communication barrier between deaf and hearing people, several sign language recognition (SLR) systems have been proposed [1]. In addition, automated expressiveness/behavior recognition from body language can also be used to reduce the major gaps that currently prevent deaf people from easily interact with hearing people.

The evaluation and validation of such automatic recognition systems rely on the availability of appropriate ground-truth data. Herein, a video-based SL and body expressiveness database, called CorSiL, is presented. It can be used for the evaluation and validation of (i) SLR systems and (ii) expressiveness/behavior recognition systems. In this regard, the CorSiL database is composed by two distinct datasets, one suitable for SLR tasks, called signLangDB, and another for expressiveness recognition from body behavior, called corpLangDB. Both datasets have been already manually annotated. By the end of the annotation task, the entire CorSiL database will be made freely available to the research community for benchmark purposes. To our knowledge it will be the first database that gathers SL videos along with videos depicting the duo-interaction between deaf and/or hearing people. This composition makes the CorSiL database so unique and valuable, since the possibility of understanding the emotions and expressiveness behind the signs may open new research paths in SLR.

2 Related Work

Several sign language databases have been proposed in the literature. A brief review on video-based SLR databases is presented in [3]. The SL datasets can be roughly classified in two main groups: for the isolated sign recognition and the continuous sign recognition. A selection of the most relevant benchmark datasets in SLR is presented below.

The Purdue RVL-SLLL ASL Corpus [4] is an available database of the American Sign Language (ASL) suitable for both isolated and continuous SLR. The database provides a wide range of signed material, including 62 isolated gestures representing the numbers and the alphabet as well as examples of short discourse narratives. The database was collected from 14 signers under controlled lightning conditions in a uniform background. The RWTH-BOSTON Corpora [2] is composed by different subsets created to be used for benchmarking of ASL recognition systems. The acquisition and recording conditions were the same in all databases. Data was collected in a dark studio background and the signer's clothing was constrained with long sleeves. The signing data were recorded using four cameras: two cameras placed towards the signer's face for stereo vision purposes, one on the side and the other in the front, with close zoom on the face. The RWTH-BOSTON-50 Corpus, created for the task of isolated SLR, contains 50 signs that were performed by three signers. On the other hand, the RWTH-BOSTON-104 database, created for continuous SLR, comprises 201 continuous sentences constructed from 104 signs. For the evaluation of hand tracking methods, this database has been annotated with the signer's hand and head positions. The larger database of the RWTH-BOSTON Corpora, called RWTH-BOSTON-400, contains a total of 843 continuous sentences created from about 400 signs, performed by 4 native signers. Both Purdue and RWTH-BOSTON databases are not suitable to signer-independent continuous sign language recognition, since the Purdue database has a small number of sentences and the RWTH-BOSTON was only performed by 4 signers. A more complete database was presented in [5], the SIGNUM database. The database currently contains 450 basic signs, representing different words types, and 780 continuous sentences. The entire corpus was performed once by 25 native signers in a controlled environment along with clothing constrains.

Although several SL databases have been proposed in the literature, many issues remain unexplored: (i) there are no Portuguese Sign Language (PSL) databases available; (ii) most of the available datasets were recorded in a very controlled scenario; (iii) there are few sign language databases that gather RGB color data with depth information; and (iv) there are no databases with videos depicting the interaction between deaf and hearing people. The database presented in this paper attempts to address all of these problems.

3 CorSiL Description

The CorSiL database was created within the framework of a research project at INESC TEC, FEUP, Porto, Portugal, which aimed to develop both a video-based automatic SLR system and a expressiveness analysis method of the human body during the interaction between a pair of subjects. In this manner the CorSiL database has two major components (or subsets) each one with a specific purpose:

1. **[signLangDB]:** a Portuguese Sign Language video dataset.
2. **[corpLangDB]:** a duo-interaction video dataset between Deaf and/or hearing people.

Besides video content, both datasets have technical annotations that are still in development. By the end of the annotation, the entire database will be made freely available for research and benchmark purposes in order to establish the first PSL database as well as the first duo-interaction database between Deaf and hearing people.

The contact with the signers and volunteers of the recordings was obtained with a partnership with the *Escola EB2/3 Eugénio de Andrade*

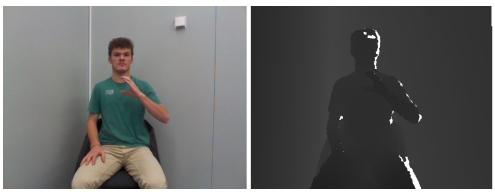

Figure 1: Color and depth pair of images from the signLangDB dataset.

and *Escola Artística de Soares dos Reis*, Porto, Portugal.

3.1 signLangDB subset

The signLangDB dataset is a PSL database suitable for both isolated and continuous SLR tasks. The dataset contains 182 isolated signs, representing the alphabet and the numbers as well as nouns, pronouns, verbs or common expressions, some performed with one hand and others with both. These signs include not only the informative part of the sign but also the entire movement from the rest position to the return to it. It also contains 40 continuous sentences that were selected in an attempt to comprise the most common situations that Deaf people might find in their daily life. All sentences are grammatically well-constructed in which there are no constraints regarding a specific sentence structure. In addition, no intentional pauses are placed between signs within a sentence.

All gestures and sentences were performed once by 15 native signers, including 5 males and 10 females, in a free and natural expression environment, without any clothing restriction but with a slightly-controlled uniform background. Moreover, some of the signers performed their gestures from a standing position while others performed seated in a chair. The recording conditions were set with this minimal amount of constrains so that they could meet a real environment scenario.

The signing data were acquired using the Microsoft Kinect camera, making this dataset one of the few with depth information associated to the RGB color data. The usage of depth eases the effort on preprocessing leaving space for heavier computational tasks. All videos were recorded using an image resolution of 640*x*480 at 30 fps. This spatial information should ensure a reliable extraction of hand and facial features from the images. Each video clip was stored as a sequence of *.png* images in order to speed up the access to individual frames. Figure 1 illustrates a pair of color and depth images.

The annotations of the signLangDB database include the segmentation of each sign and sentence for classification purposes as well as the identification of hand and head positions for tracking purposes.

3.2 corpLangDB subset

The corpLangDB dataset contains videos depicting the interaction that occur between a pair of individuals during a dialogue. The purpose of such a dataset is to enable the possibility of performing studies that analyse dialogue relationships (from sociological, psychological and technical perspectives) between two individuals, from distinct populations: Deaf and hearing people, in a relaxed environment. The conversation scenarios and topics recorded in this dataset were defined by socio-psychologists from the *Faculdade de Psicologia e Ciências da Educação da Universidade do Porto*. In this regard, the following three conversation scenarios were defined: 1) Conversation between two Deaf people; 2) Conversation between two hearing people; 3) Conversation between a Deaf and a hearing person.

In order to execute these scenarios, two requirements were defined so that the interaction between each pair of individuals could occur in the most natural way possible. Therefore, the individuals should know each other and have some kind of affinity and the acquisition should take place in a venue that was familiar to all subjects.

As the focus of the corpLangDB database is to enable the analysis of behaviour and expressiveness, the set of conversation topics was defined in a staggered way so that the discussion would generate emotions of increasing intensity in the actors of the conversation. To build a framework for the videos' acquisition, four different conversation topics were defined belonging to two-fold moments: positive (1 and 2) and negative (3 and 4): 1) Talk about happy moments; 2) Talk about people with which the actor has a strong love or friendship bond; 3) Talk about sad moments; 4) Talk about situations that awaken anger/indignation/injustice. The volunteer subjects, 13 in total, were coupled so that the conversation scenarios were covered. Each conversation between a pair of subjects was designated as a session having been a total of 9 sessions included so far in the database.

Figure 2: The same frame recorded from the cameras P0, P1, P2 and P3.

Figure 2 represents the entire scenario used to record the videos. IP0, IP1, IP2 and IP3 represent the cameras used and K a Microsoft Kinect. These were all placed in strategic locations (at a height of 2.58 meters) for the best capture possible. Two chairs were centred in the room in a way that was propitious for the dialogue in terms of proximity and comfort and for the video acquisition. The annotations available in this dataset include the positions of the head, hands, trunk, elbows, eyes, mouth, and nose. These annotations were performed using the VIPER-GT tool. [1]

4 Conclusions

This paper presents a video database, called CorSiL, that is composed by two distinct subsets, namely a SL dataset and a duo-interaction dataset (between Deaf and hearing people). The database includes technical annotations, such as temporal segmentation of the signs and body parts. The CorSiL dataset aims to provide a benchmarking tool for the comparison of both sign language recognition systems as well as expressiveness/behaviour analysis methods from body language. As future work, the technical annotations will be extended to the entire content of the database. In addition, further recordings will be conducted in order to increase the number of signers of the database.

5 Acknowledgments

This work is financed by the ERDF European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness), FCOMP-01-0124-FEDER-037281 and by National Funds through the FCT Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project PEst-C/EEI/LA0014/2013. The authors would like to thank both *Escola EB2/3 Eugénio de Andrade* and *Escola Artística de Soares dos Reis* for contacting the signers and providing the necessary means for this study. We also thank to Kelly Rodrigues and the Social Psychology Research Group of the University of Porto for their scientific advice. The fifth author would like to thank FCT for the financial support of the PhD grant with reference SFRH/BD/51430/2011.

References

[1] M.E. Al-Ahdal and N.M. Tahir. Review in sign language recognition systems. In *Computers Informatics (ISCI), IEEE Symposium on*, pages 52–57, 2012.

[2] P. Dreuw, C. Neidle, V. Athitsos, and *et al*. Benchmark databases for video-based automatic sign language recognition. *in LREC*, 2008.

[3] P. Dreuw, J. Forster, and H. Ney. Tracking benchmark databases for video-based sign language recognition. *in Trends and topics in computer vision*, 6553:286–297, 2012.

[4] AM. Martinez, R.B. Wilbur, R. Shay, and AC. Kak. Purdue rvl-slll asl database for automatic recognition of american sign language. In *Multimodal Interfaces, Fourth IEEE International Conference on*, pages 167–172, 2002.

[5] U. von Agris and K.-F. Kraiss. Towards a video corpus for signer-independent continuous sign language recognition. *Gesture in Human-Computer Interaction and Simulation*, 2007.

[1] `http://viper-toolkit.sourceforge.net/`