SIGN LANGUAGE RECOGNITION BASED ON HAND AND BODY SKELETAL DATA

Dimitrios Konstantinidis, Kosmas Dimitropoulos and Petros Daras

ITI-CERTH, 6th km Harilaou-Thermi, 57001, Thessaloniki, Greece

ABSTRACT

Sign language recognition (SLR) is a challenging, but highly important research field for several computer vision systems that attempt to facilitate the communication among the deaf and hearing impaired people. In this work, we propose an accurate and robust deep learning-based methodology for sign language recognition from video sequences. Our novel method relies on hand and body skeletal features extracted from RGB videos and, therefore, it acquires highly discriminative for gesture recognition skeletal data without the need for any additional equipment, such as data gloves, that may restrict signer's movements. Experimentation on a large publicly available sign language dataset reveals the superiority of our methodology with respect to other state of the art approaches relying solely on RGB features.

Index Terms — Sign language, deep learning, linear dynamic system, skeletal data

1. INTRODUCTION

Sign language is a structured set of hand gestures with a specific meaning that is employed from hearing impaired people in order to communicate in everyday life. Although automated SLR is very important for such people, it remains a challenging and largely unexplored research area. This is due to the fact that sign language features thousands of signs, sometimes differing only by subtle changes in hand motion, shape or position and involving significant finger overlaps and occlusions. Combined also with differences in the signing style between individuals, SLR can be very challenging for current computer vision algorithms. Finally, the unavailability of large sign language datasets and the fact that sign language is not universal but presents significant variations based on the ethnicity of signers pose challenges to the development of an accurate and robust SLR system.

Previous work on SLR can be categorized based on the data acquisition method, resulting in either direct measurement or vision-based approaches. Direct measurement methods rely on motion data acquired by data gloves, sensors or motion capturing systems [1][2]. The extracted motion data can provide accurate tracking of hands, fingers and other body parts, leading to the development of robust SLR methodologies, at the expense of costly setups and obtrusive systems as the movements of a signer are severely restricted from wearing the input devices.

On the other hand, vision-based SLR approaches rely on the extraction of discriminative spatial and temporal features from RGB images. Although unobtrusive, such methodologies present inaccuracies due to hand and finger overlaps. Most vision-based SLR methods attempt to initially track and extract hand regions prior to their classification to gestures. Hand detection has been achieved by semantic segmentation and skin color detection as skin color is usually easy to distinguish [3][4]. However, due to the fact that other body parts (e.g., face and arms) can be mistakenly recognized as hands, recent hand detection methods rely also on face detection and subtraction and background subtraction to identify only the moving parts of a scene [5][6]. To achieve accurate and robust hand tracking, especially in cases of occlusions, previous methods employ filtering techniques, such as Kalman and particle filters [6][7].

As far as hand gesture classification is concerned, the success of HMMs on several tasks, such as speech recognition has led to their use on SLR as well. Most SLR systems employ the original or modified versions of HMMs on the extracted hand motions and shapes in order to accurately detect and classify hand gestures [8][9][10]. Other successful SLR methodologies rely on distances between histograms of optical flow [5] and feature covariance matrices computed from the intensity of pixels [6], extracted from the detected hand regions.

The superb performance of deep learning techniques on several computer vision tasks has led to their adoption for SLR as well. More specifically, Koller et al. proposed a hybrid SLR system based on a convolutional neural network (CNN) and a HMM, where the CNN was employed in order to identify the hand shape and its probabilistic output was then fed to a HMM in order to guide its inference [11]. Later on, the same authors improved their SLR methodology by additionally employing bidirectional recurrent neural networks, in the form of Long Short Term Memory (LSTM) units [12].

In this work, we propose a novel methodology that bridges the gap between direct measurement and vision-based approaches thus taking advantage of both methods and overcoming their limitations. More specifically, we propose a novel SLR system that is based on the processing of video sequences in order to extract accurate body and hand skeletal data that will then be employed for robust SLR. The initial processing of the video sequences for the extraction of skeletal data is based on the works of [13][14] and their further analysis and classification is based on a proposed robust deep learning architecture. In this way, we propose a holistic SLR system that is unobtrusive as it is based only on video sequences without the need for sensors that limit the movements of signers. Moreover, our proposed system achieves accurate and robust SLR results since it relies on highly discriminative skeletal data.

The rest of the paper is organized as follows: Section 2 presents the proposed vision-based SLR methodology, while Section 3 evaluates the proposed method on a public dataset and compares it with a HMM-based SLR method. Finally, Section 4 summarizes the work of this paper by drawing conclusions and presents our plans for future work.

2. METHODOLOGY

The proposed SLR system constitutes the first attempt to merge a vision-based approach (i.e., processing of video sequences) with the accurate extraction of skeletal data without employing data gloves or other sensors that limit the movements of a signer. The architecture of the proposed SLR system is presented in Figure 1 and is extensively analyzed below.

978-1-5386-6125-3/18/$31.00 ©2018 European Union

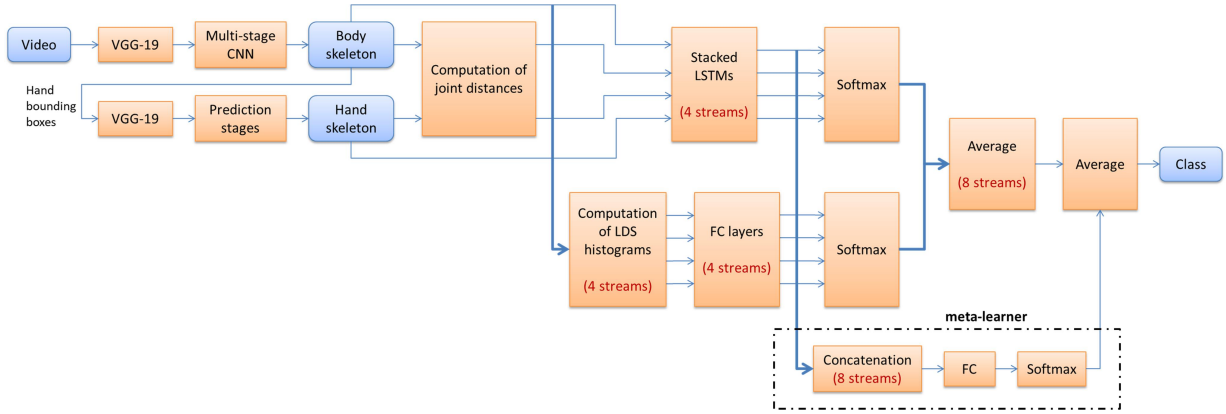

Figure 1: Proposed SLR system.

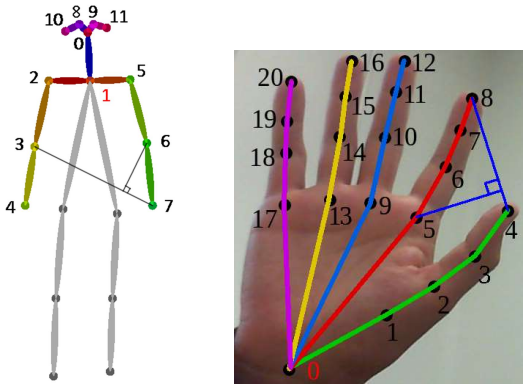

Figure 2: Body and right hand skeleton joints employed by our proposed SLR system. The red numbers correspond to the local coordinate system chosen for invariance to human position on image. Examples of how joint-line distances are computed are also presented.

The extraction of body and hand skeletal data from videos is based on the works of [13][14]. More specifically, a pre-trained on ImageNet VGG-19 network [15] up to conv4_4 is employed as feature extractor for hand skeleton detection, while the first 10 layers of the same network are employed for body skeleton detection. The outputs of the body and hand skeleton detection networks are 18 body and 21 hand 2D joints, accompanied by confidence scores. As the hand skeleton detector requires a bounding box around the hand, the wrist and elbow positions, computed from the body skeleton detector are employed in order to get an approximate position of the hand location and generate a bounding box.

In this work, we employ 12 out of the 18 extracted body skeleton joints as shown in Figure 2. This is due to the fact that i) the signers of a sign language dataset are usually sited and thus the leg skeleton joints are not visible and ii) the leg joints do not provide any valuable information for SLR tasks. Although the employed body skeleton detector produces coordinates for non-visible joints as well, their confidence score is low and thus they are deemed inappropriate by our methodology for robust classification. On the other hand, all hand joints are considered although some of these joints may be occluded by other parts of the hand and thus their confidence scores are low.

Another problem that has to be dealt with in the context of sign classification using the LSA64 dataset [10] is the fact that some of the gesture classes are signed with the right hand, while others are signed with both hands. To overcome this problem, we consider only the right hand joints for our proposed SLR system. Finally, there are also instances, where the hand skeleton detector is not able to recover the joints of the right hand in some of the frames of a video sequence. In such occasions, we employ the hand joint coordinates of the previous frames in order to fill the missing information.

Before introducing the skeletal features to the proposed skeleton classification network, a preprocessing is required. More specifically, all 2D joint coordinates are initially transformed from the image to a local coordinate system by placing the neck of the body skeleton and wrist of the hand skeleton at the origin. These origins are depicted with red color in Figure 2. The purpose of this preprocessing is to make skeleton data invariant to the absolute location of the human in the scene.

The skeleton classification network of the proposed SLR system is based on two types of spatial features and a type of temporal features. The first type of spatial features is the absolute right hand and body joint coordinates. The second type of spatial features that are employed in this work is the joint-line distances [16]. Joint-line distances model the distances from each joint to its projections on the lines formed by every other skeleton joint pair (see examples in Figure 2). Given three different joints of a skeleton $J_1, J_2, J_3 \in R^3$, the distance $d(J_1, J_2 \rightarrow J_3)$ between joint J_1, and the line formed by J_2 and J_3, is given by employing Heron's formula as follows:

$$d(J_1, J_2 \rightarrow J_3) = \frac{2\sqrt{s(s - d(J_1,J_2))(s - d(J_2,J_3))(s - d(J_3,J_1))}}{d(J_2,J_3)} \quad (1)$$

where $d(*,*)$ denotes the distance between two 2D joint coordinates and $s = 0.5 * (d(J_1,J_2) + d(J_2,J_3) + d(J_3,J_1))$. The motivation behind the selection of the joint-line distances lies in the fact that these features constitute an alternative spatial skeleton representation that models the relationship between joints. As a result, joint-line distances can complement absolute joint coordinates, forming a more descriptive spatial representation that can significantly improve SLR results. Body and hand joint coordinates and joint-line distances form a four-stream deep neural network that consists of stacked LSTM layers, having as a task to produce descriptive temporal information from the spatial features.

Other than the spatial features, linear dynamical system (LDS) histograms [17] are employed. A LDS histogram constitutes a temporal representation with the ability to capture dynamics of a multi-dimensional signal, as is the case with the joint coordinates and joint-line distances. A LDS histogram is computed by projecting sub-sequences of a signal as points to a high-dimensional space, called Grassmanian manifold [18]. In this work, we employ a spatial pyramid and compute LDS histo-

grams in each pyramid level before concatenating them in a large LDS descriptor. These descriptors are further processed with fully connected (FC) layers in order to derive more discriminative features. The resulting eight streams are finally fed to softmax layers so as each stream produces its own probabilities of a given video sequence to belong to a certain class. These probabilities are averaged and a new probability per class is produced taking into consideration all streams of the proposed skeleton classification network.

Finally, a meta-learner (see dotted outline in Figure 1) is proposed to further improve the system's accuracy. The purpose of the meta-learner is to combine the features of the eight streams by weighing them differently based on how significant their contributions are for the given SLR task. Furthermore, the meta-learner combines the features in order to produce even more discriminative ones. In this way, we enhance the learning procedure and improve the discrimination and generalization ability of the proposed SLR system. The probabilities per class computed by the meta-learner are fused (i.e., averaged) with the average class probabilities of the rest of the skeleton classification network, leading to the selection of the most probable class for a given video sequence.

3. EXPERIMENTAL EVALUATION

In this section, we initially present the tested dataset and the experimental setup before we describe the architectural details of the proposed SLR system. Then, we compare our proposed system with the methodology described in [19]. Finally, we analyze and assess the contributions of the various employed features to the performance of the proposed SLR system.

3.1 Dataset description and experimental setup

The LSA64 dataset [10] is a large Argentinian sign language dataset that consists of 10 subjects, executing 5 repetitions of a total of 64 different and commonly used types of signs. As a result, the LSA64 dataset consists of 3200 videos of different length (i.e., number of frames). In order to be employed in our proposed SLR system, all video sequences are processed so that they are composed of 48 frames each. This is achieved by employing a spline interpolation technique among the given frames of a video sequence. The experimental setup is based on [19]. More specifically, the dataset is split randomly in a training set consisting of 80% of the samples and a test set consisting of the remaining 20% of the samples. This procedure is repeated 5 times, where in each iteration, a different split of the dataset is performed. The reported results are based on the average and standard deviation of the results of all iterations.

3.2 Hyper-parameters of proposed SLR system

The optimization of the hyper-parameters that affect the performance of the proposed SLR system is performed after experimentation on the LSA64 dataset. These hyper-parameters define the size and number of LSTM and FC layers, dropout percentage, batch size and learning rate. More specifically, one- or two-layer LSTM and FC layers are considered, consisting of 128, 256, 512 or 1024 neurons, while the dropout percentage is in the range [0.0-0.5]. The streams fed with the skeleton joint coordinates employ two-layer LSTMs, while the streams fed with the skeleton joint-line distances employ one-layer LSTMs, all with optimized number of neurons and dropout percentage. On the other hand, the FC layers are all fixed with two layers consisting of 512 and 128 neurons respectively, while the FC layer of the meta-learner consists of 128 neurons. Finally, the network is implemented in Keras-Tensorflow framework and trained using the Adam optimizer with batch size of 32 and learning rate equal to 0.0001.

Table 1: Experimental evaluation on the LSA64 dataset.

Method	Accuracy(Mean ± Std)
ALL-BF-SVM [19]	95.08 ± 0.69
ALL (sequence agnostic) [19]	97.44 ± 0.59
ALL-HMM [19]	95.92 ± 0.95
Body features	93.91 ± 1.24
Hand features	91.64 ± 1.01
Deep Network without meta-learner	97.16 ± 0.57
Proposed Deep Network	**98.09 ± 0.59**

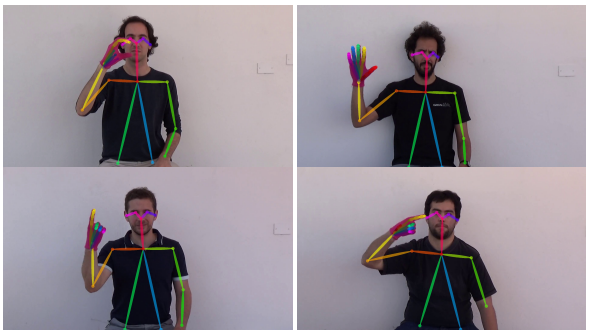

Figure 3: Extracted hand and body joints from video sequences of the LSA64 dataset.

3.3 Results

The evaluation of the various types of features in the performance of our proposed SLR system on the LSA64 dataset, along with a comparison with the SLR methods in [19] is presented on Table 1.

In [19], the authors proposed a SLR system based on the output of two classifiers, one for each hand. The classifier for each hand receives as input a sequence of cropped hand regions and normalized hand positions and employs three sub-classifiers that each use position, movement and hand-shape information. The outputs of these sub-classifiers are merged to a final probability, stating in which class a given hand gesture sequence belongs to. The authors developed their sub-classifiers in a way to be sequence agnostic, meaning that they do not rely much on the correct sequence of the hand gestures and they called their method ALL. Furthermore, they employed two more variants of their method, one of which employs HMMs with Gaussian Mixture Models output probabilities (ALL-HMM) and the other transforms their features to binary ones and then employs one-versus-all multi-class Support Vector Machines (ALL-BF-SVM).

From Table 1, a few conclusions can be drawn. Firstly, the body features constitute a slightly better representation than the hand features for sign language recognition since they achieve a 2.27% increase in sign language recognition on the LSA64 dataset. This is attributed to the fact that the body joints are more reliably and robustly detected than the hand joints. Accurate hand joint detection suffers from occlusions and overlaps between the fingers and as a result, no detector can reliably infer the locations of non-visible joints. This can also be observed by the low confidence scores the employed hand skeleton detector produces.

However, the employment of both hand and body skeletal features is beneficial for the SLR task. Hand skeletal joints contain valuable knowledge that can complement the information body skeletal joints provide and therefore, their combined use gives a boost to the performance of our proposed SLR deep net-

work as shown by the increase of 3.25% in the accuracy achieved on LSA64 dataset. It is also worth mentioning that although our proposed SLR system does not employ any information from the left hand, our methodology manages to successfully classify both one-handed and two-handed signs of the LSA64 dataset, demonstrating the discrimination power of the employed features.

Moreover, the use of a meta-learner is beneficial to the performance of our proposed SLR methodology. This can be attributed to the construction of highly discriminative features by the employed meta-learner based on the corresponding features each data stream produces. In this way, the meta-learner exploits the derived meta-knowledge, enhances the learning procedure and improves the discrimination and generalization ability of the proposed SLR system. Finally, our proposed methodology outperforms all variants of the SLR method of [19] by at least 0.65% no matter what types of classifiers or features they employed. Finally, examples of the detected hand and body joints that are extracted from the video sequences of the LSA64 dataset and are employed for the classification of sign language gestures are shown in Figure 3.

4. CONCLUSIONS AND FUTURE WORK

Previous works on sign language recognition were based either on direct measurement of skeletal data from obtrusive sensors and data gloves or inaccurate processing of video sequences. This paper presents a novel SLR system that attempts to overcome the limitations of previous methods by proposing the extraction and processing of hand and body skeletal data from video sequences. The experimentation on LSA64 dataset shows that our SLR system outperforms other vision-based SLR approaches, despite difficulties in extracting accurate skeletal data due to occlusions. As a future work, we plan to test our novel SLR system in additional sign language datasets and study the contribution of image and optical flow features in the task of sign language recognition.

5. ACKNOWLEDGEMENT

This work has been supported from EC under grant agreement no. H2020-ICT-19-2016-2 “EasyTV: Easing the access of Europeans with disabilities to converging media and content”.

6. REFERENCES

[1] G. Fang, W. Gao and D. Zhao, “Large vocabulary sign language recognition based on fuzzy decision trees,” *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 34, no. 3, pp. 305-314, May 2004.

[2] W.W. Kong and S. Ranganath, “Signing Exact English (SEE): Modeling and recognition,” *Pattern Recognition*, vol. 41, no. 5, pp. 1638-1652, 2008.

[3] S.S. Rautaray and A. Agrawal, “A Real Time Hand Tracking System for Interactive Applications,” in *International Journal of Computer Applications*, vol. 18, no. 6, pp. 28-33, March 2011.

[4] Z. Zhang and F. Huang, “Hand Tracking Algorithm Based on SuperPixels Feature,” in *International Conference on Information Science and Cloud Computing Companion*, Guangzhou, December 2013, pp. 629-634.

[5] K.M. Lim, A.W.C. Tan and S.C. Tan, “Block-based histogram of optical flow for isolated sign language recognition,” *Journal of Visual Communication and Image Representation*, vol. 40, part B, pp. 538-545, 2016.

[6] K.M. Lim, A.W.C. Tan and S.C. Tan, “A feature covariance matrix with serial particle filter for isolated sign language recognition,” *Expert Systems with Applications*, vol. 54, pp. 208-218, 2016.

[7] Y.F.A. Gaus and F. Wong, “Hidden Markov Model-Based Gesture Recognition with Overlapping Hand-Head/Hand-Hand Estimated Using Kalman Filter,” in *Third International Conference on Intelligent Systems Modelling and Simulation*, Kota Kinabalu, 2012, pp. 262-267.

[8] N. Tanibata, N. Shimada and Y. Shirai, “Extraction of Hand Features for Recognition of Sign Language Words,” in *International Conference on Vision Interface*, 2002, pp. 391-398.

[9] X. Ni, G. Ding, X. Ni, X. Ni, Q. Jing, J. Ma, P. Li and T. Huang, “Signer-Independent Sign Language Recognition Based on Manifold and Discriminative Training,” in *Information Computing and Applications*, 2013, pp. 263-272, Springer Berlin Heidelberg.

[10] F. Ronchetti, F. Quiroga, C. Estrebou, L. Lanzarini and A. Rosete, “LSA64: A Dataset of Argentinian Sign Language,” *XX II Congreso Argentino de Ciencias de la Computación (CACIC)*, 2016.

[11] O. Koller, S. Zargaran, H. Ney, and R. Bowden, “Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition,” in *Proceedings of British Machine Vision Conference (BMVC)*, 2016.

[12] O. Koller, S. Zargaran and H. Ney, “Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 3416-3424.

[13] T. Simon, H. Joo, I. Matthews and Y. Sheikh, “Hand Keypoint Detection in Single Images Using Multiview Bootstrapping,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 4645-4653.

[14] Z. Cao, T. Simon, S.-E. Wei and Y. Sheikh, “Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1302-1310.

[15] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *International Conference on Learning Representations (ICLR)*, 2015.

[16] S. Zhang, X. Liu and J. Xiao, “On Geometric Features for Skeleton-Based Action Recognition Using Multilayer LSTM Networks,” in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2017, pp.148-157.

[17] K. Dimitropoulos, P. Barmpoutis and N. Grammalidis, “Higher order linear dynamical systems for smoke detection in video surveillance applications,” *IEEE Transactions on Circuits and Systems for Video Technology* vol. 27, no. 5 pp. 1143-1154, 2017.

[18] K. Dimitropoulos, P. Barmpoutis, A. Kitsikidis and N. Grammalidis, “Classification of Multidimensional Time-Evolving Data using Histograms of Grassmannian Points,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 4, pp. 892-905, April 2018.

[19] F. Ronchetti, F. Quiroga, C. Estrebou, L. Lanzarini and A. Rosete, “Sign Language Recognition Without Frame-Sequencing Constraints: A Proof of Concept on the Argentinian Sign Language,” in *Advances in Artificial Intelligence - IBERAMIA 2016*, 2016, pp. 338-349, Springer International Publishing.