A Deep Learning Approach for Analyzing Video and Skeletal Features in Sign Language Recognition

Dimitrios Konstantinidis
Information Technologies Institute
CERTH
Thessaloniki, Greece
dikonsta@iti.gr

Kosmas Dimitropoulos
Information Technologies Institute
CERTH
Thessaloniki, Greece
dimitrop@iti.gr

Petros Daras
Information Technologies Institute
CERTH
Thessaloniki, Greece
daras@iti.gr

***Abstract*—Sign language recognition (SLR) refers to the classification of signs with a specific meaning performed by the deaf and/or hearing-impaired people in their everyday communication. In this work, we propose a deep learning based framework, in which we examine and analyze the contribution of video (image and optical flow) and skeletal (body, hand and face) features in the challenging task of isolated SLR, in which each signed video corresponds to a single word. Moreover, we employ various fusion schemes in order to identify the optimal way to combine the information obtained from the various feature representations and propose a robust SLR methodology. Our experimentation on two sign language datasets and the comparison with state-of-the-art SLR methods reveals the superiority of optimally combining skeletal and video features for SLR tasks.**

Keywords—sign language, deep learning, optical flow, skeletal data, probability fusion

I. Introduction

Variants of sign languages are employed by millions of deaf or hearing-impaired people around the world in order to communicate in their everyday life. As a result, the automatic translation of sign languages is considered vital and important for these people. Unfortunately, there are several reasons that render SLR a challenging research area. It is true that sign language differs from country to country, making the creation of a universal SLR system infeasible. Furthermore, the significant sign language variations based on the ethnicity of signers pose challenges to the creation of a large publically available dataset. This means that there is not a widely acceptable sign language dataset for experimental evaluation and most SLR methods are evaluated on their own small and usually inadequate datasets.

Other problems that can increase the difficulty of developing an accurate and robust SLR methodology are the variations in the signing style between individuals, as well as, the variations in the signing style of the same individual. Additionally, each sign language consists of thousands of signs that can differ by subtle changes in hand shape, motion and position. Finally, the capturing of the most important body parts in sign language, i.e., hands, suffers from significant finger overlaps and occlusions.

A key challenge in SLR is the design and extraction of visual descriptors that can reliably identify and classify signs. To this end, we propose an isolated SLR system that extracts discriminative features from videos. More specifically, we initially extract video (image and optical flow) and skeletal (body, hand and face) features from videos, then derive more descriptive ones by employing simple recurrent deep learning modules and finally fuse the information of the deep modules by means of a meta-learner and a probability fusion scheme. To this end, we also compare and evaluate the contribution of various fusion schemes to the performance of our proposed SLR methodology. The contributions of this work are summarized below:

a) Our proposed method comprises the first attempt to combine video and skeletal features in a holistic SLR system based on a deep learning approach.

b) Our proposed SLR system combines different streams of information related to the motion of a singer, such as body, hand and face features.

c) We evaluate various fusion schemes in an attempt to optimally combine the information from the various data streams.

The rest of the paper is organized as follows: in Section II, we perform a literature review of state-of-the-art SLR methods. In Section III, we present our proposed methodology for the extraction and classification of video and skeletal features, as well as the various proposed fusion schemes. Finally, in Section IV, we perform an experimental evaluation and comparison of the proposed method against other SLR techniques, while in Section V, we summarize our work by drawing conclusions.

II. Related Work

The criterion of data acquisition method is often used in the literature to classify SLR methodologies in two fundamental categories. The first category consists of direct measurement methods that involve the use of data gloves, multiple sensors attached to hands and body and motion capturing systems [1][2]. Such sensors allow the extraction of accurate motion data that describe the movement of hands, head and other body parts that are useful for the classification of signs. As a result, direct measurement methods can lead to the development of accurate and robust SLR methodologies, at the expense of complicated and costly setups and obtrusive systems because the movements of a signer are severely restricted from being in direct contact with the input devices.

The second category of SLR methods consists of vision-based approaches. These approaches attempt to overcome the limitations of the direct measurement methods by relying on the extraction of descriptive temporal and spatial features from videos captured by cameras and depth sensors [3][4]. Although the vision-based methods allow a signer to perform gestures with no restrictions, the data gathered can be noisy and inaccurate due to overlaps and occlusions between fingers and other body parts. Next, we present a review of previous works related to vision-based approaches since our proposed methodology can be classified as such.

In order to classify hand gestures, most vision-based SLR approaches attempt to initially detect and extract hand

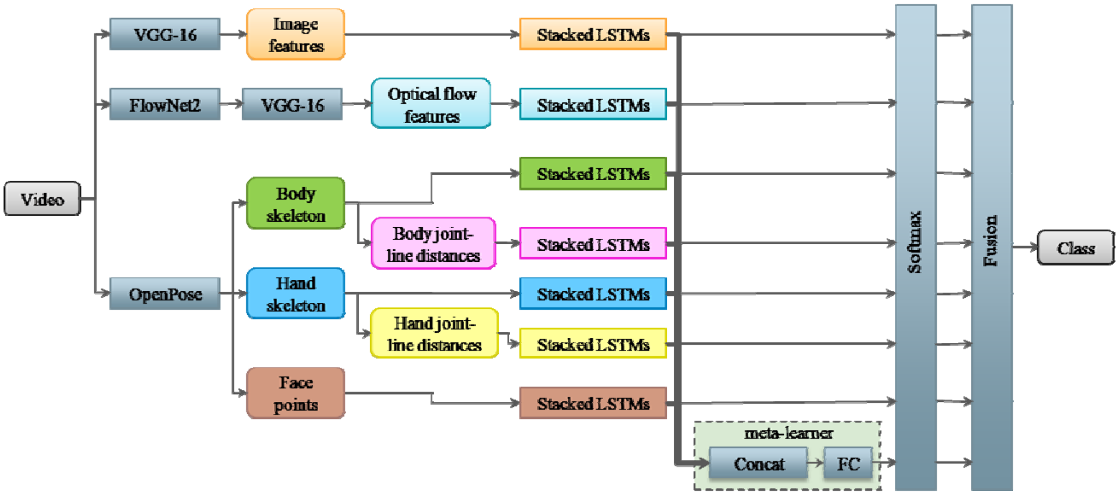

Figure 1: Block diagram of the proposed SLR methodology. Each data stream is shown with a different color.

regions. Traditionally, hand detection is achieved by skin color detection and semantic segmentation [5][6]. Unfortunately, other body parts, such as face and arms possess similar skin color information; thus they can be erroneously recognized and extracted along with hands. As a result, recent hand detection methods rely also on face detection and background subtraction in order to identify only the moving parts of a scene [7][8]. Moreover, hand detection methods usually employ tracking techniques, such as Kalman and particle filters in order to handle occlusion problems and achieve accurate and robust hand detection and extraction [8][9].

As far as hand gesture classification is concerned, several methods employ the extracted hand regions and compute distances between histograms of optical flow [7] or feature covariance matrices from pixel intensities [8]. The success of Hidden Markov Models (HMM) on several tasks, such as speech and handwriting recognition, has led to their use on SLR as well. Several SLR methods employ the original or modified versions of HMMs on the extracted hand shapes and positions in order to reliably classify hand gestures [10][11][12].

Recently, the superb performance of deep learning algorithms on several computer vision tasks has led many researchers to adopt them for SLR as well. More specifically, Koller et al. proposed a hybrid SLR system based on a convolutional neural network (CNN) and an 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 [13]. The same authors subsequently improved their SLR method by additionally employing bidirectional recurrent neural networks, in the form of Long Short Term Memory (LSTM) units [14]. On the other hand, Huang et al. proposed the use of 3D CNNs that can automatically capture both temporal and spatial information from the raw video sequences, without the need for designing features [15]. Finally, Konstantinidis et al. in [16] proposed a skeleton-based approach using recurrent neural networks and linear dynamical systems [17] so as to accurately classify signs.

In this work, we propose a novel SLR 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 system that processes video sequences in order to extract video and skeletal features that will then be employed for robust SLR. Apart from the body and hand skeletal data, in this paper, we analyze the effect of not only image and optical flow features, but also face features and investigate alternative fusion schemes in order to identify the optimal one that allows for the reliable detection and classification of gestures.

In this way, we present a holistic, unobtrusive SLR system solely based on the processing of video sequences without the need for sensors that limit the movement of signers. Finally, we show that the proposed system can achieve accurate and robust SLR results by optimally fusing highly discriminative video and skeletal features.

III. Methodology

A block diagram of the proposed SLR methodology is presented in Fig. 1, where it is shown that our method is based solely on the processing of video sequences and the extraction and fusion of discriminative features for the classification of video sequences to isolated signs.

The proposed methodology relies on the extraction of video (i.e., image and optical flow) and skeletal (i.e., body, hand and face) features from video sequences. To derive video features, we employ the VGG-16 network [18], pre-trained on ImageNet on both the raw video sequences (i.e., image features) and the optical flow images (i.e., optical flow features). In order to obtain the optical flow images, we employ the well-known and accurate optical flow deep network, FlowNet2 [19]. Regarding the skeletal features, we employ the OpenPose algorithm [20], which is a deep network, capable of producing hand and body skeleton joints and face points by processing raw videos. The positions of the provided by OpenPose skeletal data are presented in Fig. 2. The output of OpenPose is 18 body and 21 hand 2D joints and 69 2D face points. Subsequently, we discard 6 body joints, as shown in Fig. 2a, because firstly a signer is usually sited and thus the leg joints are not visible and secondly the leg joints do not participate in the signing process and thus they do not carry any valuable information.

Furthermore, only the right hand skeleton joints are extracted as the right hand is the main signing hand in our datasets, although there are signs executed by both hands.

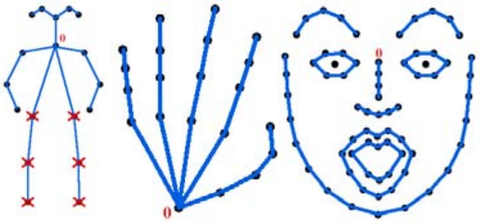
(a) (b) (c)

Figure 2: Body and hand skeleton joints and face points extracted from OpenPose [20]. We denote with red zeros the joints chosen as origins of the local coordinate systems and with red crosses the joints that are not taken into account in our proposed SLR method.

Finally, before employing the skeletal data, we normalize their positions by transforming them from image to local coordinate systems. The origins of the local coordinate systems are chosen to be the neck, wrist and upper nose point for the body, hand and face skeleton, respectively (see Fig. 2). The purpose of this transformation is to make the skeletal data invariant to the absolute location of the signer in the scene.

Two additional spatial features are employed by processing the body and hand skeletons and computing the joint-line distances [21]. Joint-line distances model the distances from each joint to its projections on the lines formed by every other skeleton joint pair. Given three different joints of a skeleton $J_1, J_2, J_3 \in R^3$, the distance $d(J_1, J_2 \to J_3)$ between joint J_1, and the line formed by J_2 and J_3, is given by employing the Heron's formula as follows:

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

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 on the fact that these features constitute an alternative spatial feature representation that models the relationship between joints. As a result, joint-line distances can complement the other feature representations, forming an additional descriptive spatial representation that can significantly improve SLR results. The reason why joint-line distances are not computed for the face points as well lies on the fact that the number of computed face points is quite large, thus leading to an enormous amount of face joint-line distances (i.e., over 150k distances). As a result, the computational complexity for the processing and classification of such enormous vectors would be really high.

Therefore, seven data streams are created from the processing of the raw video sequences. These streams are fed to stacked LSTMs, which are several LSTM units put one after the other. These units are individually optimized to achieve best performance for the SLR problem at hand. A meta-learner is also employed that concatenates the features computed from the stacked LSTMs and then processes them a bit further to derive even more powerful and discriminative features by using a fully connected (FC) layer. The purpose of the meta-learner is to combine the features of all streams by weighing them differently based on how significant their contributions are for the given SLR task. In this way, we enhance the learning procedure and improve the discrimination and generalization ability of the proposed SLR system. The seven data streams, along with the meta-learner stream, form a set of eight classifiers that are then fed to softmax layers so as each of these classifiers produces its own probabilities that a given video sequence belong to a certain class.

To fuse the aforementioned probabilities, the work in [16] proposed the averaging of the data streams, the computation of an overall probability and then again, the averaging of this probability with the probability of the meta-learner in order to obtain the final probability per class. In this work, we not only employ the aforementioned averaging fusion scheme, which we name AV in short notation, but we also investigate other fusion schemes so as to find the optimal way to combine the eight streams and improve the performance of the proposed SLR methodology.

To this end, we test majority voting (MV), which is a well-known technique that accepts as class of a tested video sequence, the one with the most votes from the employed classifiers. Furthermore, inspired by the work of [22], we employ Dynamic Score Combination (DSC) [23][24] and Particle Swarm Optimization (PSO) [25]. DSC attempts to combine the individual probabilities in a way that the combined probability distribution exhibits a larger separation than the probability distribution produced by the individual classifiers. Since DSC is employed in two-class optimization problems and our problem is multi-class, we had to adapt DSC slightly so that each class is compared against all other classes. Given the probability of a classifier for a single class p^i, with $i = 1..8$, the overall probability based on DSC is given by:

$$p_{DSC} = (1-\beta)\min_i\{p^i\} + \beta\max_i\{p^i\} \quad (2)$$

where β is the combination weight, defined by the mean rule:

$$\beta = \frac{1}{m}\left(\sum_{i=1}^{m} p^i\right) \quad (3)$$

and m is equal to the number of classifiers (i.e., $m = 8$ in our case). Another fusion scheme is the PSO, which is a global optimization algorithm, motivated by social behavior of organisms such as bird flocking and fish schooling. In this context, the overall probability of a class is given by the weighted aggregation of the individual probabilities as:

$$p_{PSO} = \sum_{i=1}^{m} w_i p^i \quad (4)$$

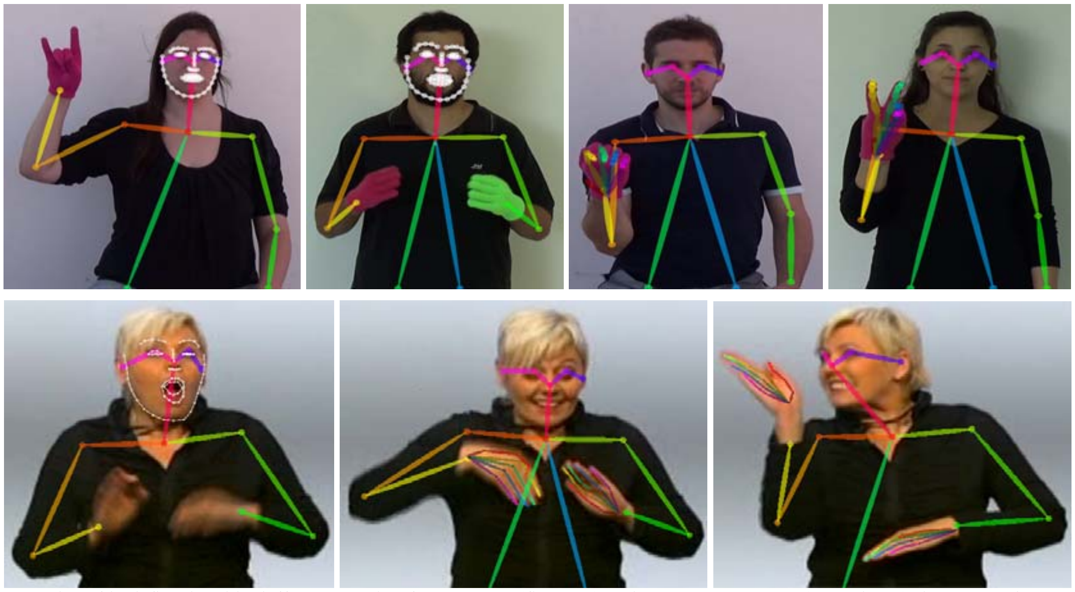

Figure 3: Examples of body/hand and body/face extraction from LSA64 (first row) and RWTH-PHOENIX (second row) datasets as the computational complexity of detecting all three skeletal features simultaneously was too heavy for our setup.

The PSO algorithm considers each single solution w_i, with $i = 1..8$, as a particle in the search space and associates this particle with a fitness value and velocity, which direct the movement of the particle. In each iteration, the algorithm tries to improve a candidate solution with regard to a given measure of quality, i.e., the fitness function to be optimized, while the particles move in the problem space by following the current optimum particle. In our case, the fitness function is the accuracy of a solution w_i, when applied on the training set of a given dataset.

Finally, we propose another fusion scheme that we call deep weight averaging (DWA). This scheme attempts to optimize a set of weights, exactly like in (4), but instead of a PSO algorithm, a deep network is employed that accepts as inputs the individual probabilities and learns a set of weights that can optimize the overall probability.

IV. Experimental evaluation

In this section, we describe the sign language datasets used for evaluation and the experiments that led to the optimal performance of our SLR methodology. Finally, we present a comparative study with other state-of-the-art SLR techniques.

A. Dataset description

Two datasets are selected for the experimental evaluation. The first dataset, called LSA64 [12], is a large Argentinean sign language dataset that consists of 10 subjects, executing 5 repetitions of a total of 64 different types of signs. As a result, the LSA64 dataset comprises 3200 videos of different length (i.e., number of frames). For the experimental evaluation of the 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 second dataset that is employed in this work is the RWTH-PHOENIX-Weather database [26], noted shortly as RWTH-PHOENIX below. This dataset was basically created for continuous SLR. However, there is a part of the dataset, called Signer03 Cut-out Gloss Recognition that allows for experiments in the context of isolated SLR. We take advantage of this setup, but we process the dataset by discarding a few samples in order to be more suitable for deep learning training. More specifically, we discard classes with fewer than 10 samples and we also discard samples from classes with over 50 samples. Furthermore, we process all video sequences in order to consist of 10 frames each. As a result, the final processed dataset that we employ consists of 50 classes with 10-50 samples per class and 1297 and 238 training and test video sequences respectively.

The reason behind the selection of these datasets lies in their special characteristics for a deep learning training. The LSA64 dataset is a large and balanced sign language dataset with several frames per video sequence and thus it is suitable for a deep learning framework. By utilizing this dataset, we want to unravel the full potential of a deep network. On the other hand, despite our changes, the second dataset remains a highly unbalanced dataset with few frames per video sequence and cases where some of these frames are blurry. As a result, the second dataset is quite challenging for a deep learning algorithm and by using it we want to observe how well our proposed SLR system can cope with problematic datasets.

B. Experimental setup

The experimental setup for the LSA64 dataset is based on [27]. More specifically, the dataset is split randomly in a training set consisting of 80% of all 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. As far as the RWTH-PHOENIX dataset is concerned, the video sequences are already split in training and test sets. What we do, in this case, is repeating the training of our proposed SLR method for 5 times, where the weights of the deep network are randomly re-initialized after each repetition. The reported results are based on the average and standard deviation of the results of all repetitions.

C. Hyper-parameter optimization

The optimization of the hyper-parameters that affect the performance of the proposed SLR method is performed after experimentation on the training sets of the LSA64 and RWTH-PHOENIX datasets. These hyper-parameters define the size and number of stacked LSTM units, size of the FC layer, dropout percentage, batch size and learning rate. More specifically, one- or two-layer LSTMs are considered, consisting of 128, 256, 512 or 1024 neurons, while the dropout percentage is in the range [0.0-0.5]. Similarly, the size of the FC layer is selected after experimentation among the values of 128, 256, 512 and 1024. We end up with 128 neurons for the FC layer, while the size, dropout and number of LSTM units vary significantly among the data streams of the proposed SLR method. Furthermore, for the image and optical flow features, we get the output of the last layer of the VGG-16 network, which is a 1024-element vector. 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.

D. Evaluation of features and fusion schemes

Here, we evaluate the individual feature representations, the meta-learner and the proposed fusion schemes in order to identify the optimal way of combining the information from the different data streams and the meta-learner of our proposed SLR methodology. The results from the experimental evaluation are presented in Table I.

From Table I, it can be deduced that the most discriminative features are the image and the optical flow features, revealing the power of the pre-trained VGG-16. Furthermore, the joint-line distances seem to constitute more powerful representations than the raw skeleton joints, thus justifying our choice to employ them for our proposed SLR methodology. On the other hand, it can be observed that the face features perform poorly in the SLR task. This is something to be expected as the face features alone are not adequate enough to differentiate signs. Their purpose is mostly complementary in order to enhance the performance of other more descriptive feature representations, such as hand and body skeletal data. Finally, the meta-learner is successful in its task of combining the various data streams in an attempt to produce even more powerful features. This can be more clearly observed in the RWTH-PHOENIX dataset, where the features that the meta-learner provides outperform all individual feature representations.

The evaluation of the proposed fusion schemes cannot give a clear view of the optimal one. From Table I, one can observe that the AV fusion scheme that was proposed in [16] under-performs with respect to the other fusion schemes. On the other hand, MV has the limitation that it does not take into account the accuracy of the individual classifiers. As a result, although it performs quite well in the LSA64 dataset because all classifiers are really accurate, it performs relatively poorly in the RWTH-PHOENIX dataset, where all classifiers have mediocre performance. The DSC fusion scheme performs optimally in the LSA64 dataset, while our proposed DWA method performs optimally in the RWTH-PHOENIX dataset. This shows the power of deep learning in not only producing discriminative features, but also weighing features appropriately in order to lead to improved results. It is also worth noting that the PSO algorithm is quite sensitive to its initialization and therefore, we executed it 5 times and obtained its mean accuracy.

Table I: Results of individual feature representations and fusion schemes.

Feature	**Dataset results (mean ± std)**	
	LSA64	**RWTH-PHOENIX**
Image	**99.37 ± 0.25**	59.66 ± 1.39
Optical flow	98.81 ± 0.49	39.24 ± 2.11
Body skeleton	91.06 ± 1.09	33.28 ± 2.48
Body joint-line distances	93.34 ± 2.23	42.52 ± 2.82
Hand skeleton	85.88 ± 1.48	29.58 ± 2.27
Hand joint-line distances	95.19 ± 0.36	44.79 ± 1.5
Face	18.22 ± 1.54	19.66 ± 1.77
Meta-learner	97.94 ± 1.03	**60.76 ± 3.21**
Fusion		
AV	99.19 ± 0.47	64.29 ± 2.93
MV	99.81 ± 0.17	64.87 ± 1.8
DSC	**99.84 ± 0.19**	67.98 ± 1.86
PSO	99.8 ± 0.06	66.49 ± 2.32
DWA	99.72 ± 0.26	**69.33 ± 1.57**

E. Comparison with state-of-the-art SLR methods

Although the RWTH-PHOENIX dataset has been evaluated in the context of continous SLR, no isolated SLR method has been applied yet. This reveals again one of the problems of SLR, which is the unavailability of significant experimental evaluation on the same dataset. To overcome this problem, we also test the deep network presented in [16] on the RWTH-PHOENIX. In Table II and Table III, our proposed SLR method is evaluated against other state-of-the-art methodologies on the LSA64 and RWTH-PHOENIX datasets respectively.

From Table II, it can be deduced that our proposed SLR method significantly outperforms all other state-of-the-art methodologies in the LSA64 dataset, reaching an almost perfect accuracy. This reveals the power of employing several alternative feature representations and a reliable fusion scheme that can boost the performance of a classification procedure even further. Similar conclusions can be drawn from Table III, in which we observe that the use of additional features and a more appropriate fusion scheme is beneficial to the performance of a SLR algorithm.

A comparison of the performance of our proposed SLR method between the two datasets can also be performed. One can observe that our method reaches an almost perfect accuracy in a balanced and large dataset (i.e.,

Table II: Experimental evaluation on the LSA64 dataset.

Method	**Accuracy (mean ± std)**
ALL-BF-SVM [27]	95.08 ± 0.69
ALL (sequence agnostic) [27]	97.44 ± 0.59
ALL-HMM [27]	95.92 ± 0.95
Deep network [16]	98.09 ± 0.59
Proposed SLR method	**99.84 ± 0.19**

Table III: Experimental evaluation on the RWTH-PHOENIX dataset.

Method	Accuracy (mean ± std)
Deep network [16]	56.13 ± 2.33
Proposed SLR method	**69.33 ± 1.57**

LSA64 dataset), but it achieves mediocre performance in a problematic dataset, such as the RWTH-PHOENIX. However, although the RWTH-PHOENIX dataset is not so suitable for a deep network, our SLR method does a fine job classifying it, achieving almost 70% accuracy among 50 classes. This fact reveals the discrimination power of the proposed feature representations, meta-learner and fusion schemes and it is the main reason we chose to test our SLR method in such a dataset.

V. Conclusions

In this paper, we propose a novel SLR method that is based on the processing of raw video sequences and the extraction of discriminative video and skeletal features. These features are then processed in a deep learning framework and classified by utilizing the power of a meta-learner and proposed fusion schemes. The results on two SLR datasets show that our proposed method can outperform other state-of-the-art methodologies, while performing robustly on unbalanced and problematic datasets. As a future work, we will concentrate on the creation of a new sign language dataset that will be suitable for a deep learning framework.

Acknowledgment

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".

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.G.M. Almeida, F.G. Guimarães and J.A. Ramírez, "Feature extraction in Brazilian sign language recognition based on phonological structure and using RGB-D sensors", Expert Systems with Applications, vol. 41, no. 16, pp. 7259-7271, 2014.

[4] C. Sun, T. Zhang and C. Xu, "Latent support vector machine modeling for sign language recognition with Kinect", ACM Transactions on Intelligent Systems and Technology (TIST), vol. 6, no. 2, pp. 1-20, 2015.

[5] 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.

[6] 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.

[7] 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 Repre-sentation, vol. 40, part B, pp. 538-545, 2016.

[8] 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.

[9] 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.

[10] 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.

[11] 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 Infor-mation Computing and Applications, 2013, pp. 263-272, Springer Berlin Heidelberg.

[12] 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.

[13] 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.

[14] 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.

[15] J. Huang, W. Zhou, H. Li and W. Li, "Sign language recognition using 3d convolutional neural networks", in IEEE International Conference on Multimedia and Expo (ICME), 2015, pp. 1-6.

[16] D. Konstantinidis, K. Dimitropoulos and P. Daras, "Sign language recognition based on hand and body skeletal data", in 3DTV conference, May 2018.

[17] 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, 2017.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", in Proc. International Conference on Learning Representations, 2014.

[19] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[20] 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.

[21] S. Zhang, X. Liu and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer LSTM networks," in IEEE Winter Conference on Applica-tions of Computer Vision (WACV), March 2017, pp.148-157.

[22] 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.

[23] L. Piras, R. Tronci and G. Giacinto, "Diversity in ensembles of codebooks for visual concept detection," in International Conference on Image Analysis and Processing (ICIAP), Naples, Italy, 2013.

[24] R. Tronci, G. Giacinto, and F. Roli, "Dynamic score combination: A supervised and unsupervised score combination method," in Machine Learning and Data Mining in Pattern Recognition (Lecture Notes in Computer Science), Springer Berlin Heidelberg, vol. 5632, 2009, pp. 163-177.

[25] S. K. Tchomté and M. Gourgand, "Particle swarm optimization: A study of particle displacement for solving continuous and combinatorial optimization problems," International Journal of Production Economics, vol. 121, no. 1, pp. 57-67, 2009.

[26] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. Piater and H. Ney, "RWTH-PHOENIX-Weather: A large vocabulary sign language recognition and translation corpus", in Language Resources and Evaluation (LREC), 2012, pp. 3785-3789.

[27] 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.