TRANSFER LEARNING FOR VIDEOS: FROM ACTION RECOGNITION TO SIGN LANGUAGE RECOGNITION

Noha Sarhan, Simone Frintrop

Universität Hamburg
Department of Computer Science
Vogt-Kölln Strasse 30, 22527 Hamburg

ABSTRACT

In this paper, we propose using Inflated 3D (I3D) Convolutional Neural Networks for large-scale signer-independent sign language recognition (SLR). Unlike other recent methods, our method relies only on RGB video data and does not require other modalities such as depth. This is beneficial for many applications in which depth data is not available. We show that transferring spatiotemporal features from a large-scale action recognition dataset is highly valuable to the training for SLR. Based on an architecture for action recognition [1], we use two-stream I3D ConvNets operating on RGB and optical flow images. Our method is evaluated on the ChaLearn249 Isolated Gesture Recognition dataset and clearly outperforms other state-of-the-art RGB-based methods.

Index Terms— Sign language recognition, video transfer learning, 3D CNNs

1 Introduction

Sign language serves as the primary means of communication amongst the deaf and hard-of-hearing. It is made up of structured sets of hand gestures governed by grammatical and contextual rules. Despite common belief, sign languages go beyond mere hand gestures; different body postures, mouth shapes, and eye gaze, as well as relative hand position contribute to deliver the complete meaning of the gesture [2]. Sign languages are not international; different countries have their own language, where similar gestures could have different meanings depending on cultural differences [3]. Having an automatic method that is able to reliably recognize sign language gestures would have a significant impact in breaking the communication barrier between speakers and non-speakers of sign language.

Although several methods have been proposed for sign language recognition (SLR), many challenges remain associated with it that affect the recognition accuracy. This is mainly attributed to the large intra-class ambiguity between gestures performed by different signers. The large number of classes, which may blow up to become as large as the dictionary size of that sign language, complicates matters further as it increases the inter-class similarity between gestures. In addition, other factors such as motion blur, performer's speed, out-of-vocabulary gestures, and various view-points should be irrelevant to the recognition of the gesture.

Fig. 1. Each row represents a sample video corresponding to one isolated gesture from ChaLearn249 IsoGD dataset [4].

With motion being an important cue in SLR, learning spatiotemporal features becomes key. It is hardly surprising that current state-of-the-art approaches for SLR are based on deep neural networks [5, 6, 7, 8, 9, 10, 11] in order to learn these features. However, when it comes to sign language, acquiring enough annotated data for training deep networks is a difficult task. It is not only tedious and time-consuming, but it also suffers from the additional problem that only experts who speak the respective sign language can perform the labelling, which makes crowd sourcing extremely difficult, if not impossible. In addition, SLR with mere RGB images is a difficult task. Researches, therefore, quickly resort to methods that rely on other modalities, such modalities include depth [8, 6, 12, 13] and/or joint locations [3]. However, depth data is not available for many applications, e.g. YouTube videos, TV broadcasts, etc. For such applications, it is mandatory to have a method that performs well relying only on RGB videos.

In this paper, we present an RGB-only method to recognize isolated sign language gestures using inflated two-stream 3D ConvNet. Our approach is based on the architecture of [1], which was designed for action recognition operating on an

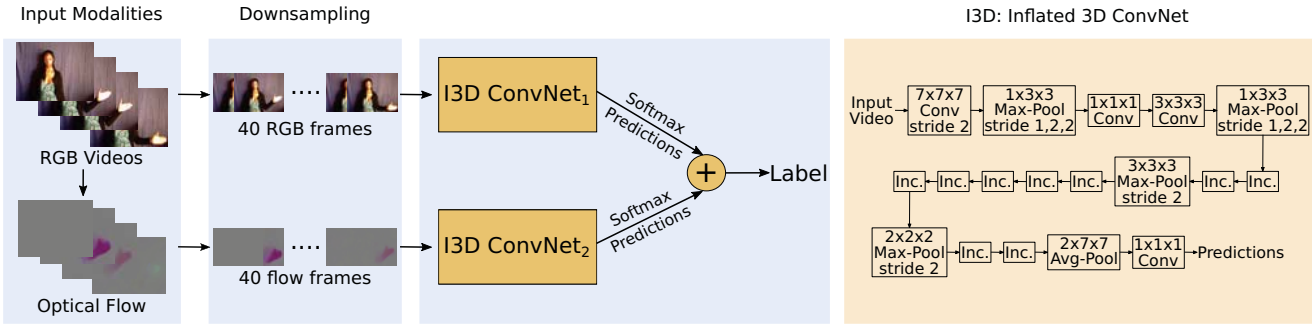

Fig. 2. Left: Pipeline of our proposed approach. RGB and optical flow data are used to train two separate inflated 3D ConvNets. The predictions of each stream are averaged during evaluation to give the final label. Right: Details of the inflated 3D ConvNets, adapted from [1]. The "Inc." blocks represent 3D Inception modules.

RGB and an optical flow stream. The network is also pre-trained on Kinetics [14], a large-scale action recognition dataset with 600 classes, each with at least 600 video clips. This enables our network to profit from the learned 3D features, which are then fine-tuned to a SLR-specific dataset, namely ChaLearn249 IsoGD [4]. We show a sample of the dataset in Fig. 1. While the task of SLR is significantly more difficult than typical action recognition, where a certain object might significantly help or even suffice for classification [9, 13], we show that this type of pre-training is valuable to SLR.

Our contributions can be summarized as follows:

- We present a simple yet effective approach for the recognition of isolated sign language gestures on RGB-only data which outperforms, to the best of our knowledge, all current state-of-the-art methods.
- We show the benefit of using pre-trained weights from action recognition tasks in SLR. To the best of our knowledge, this has until now not been applied to comparable tasks of recognition of well-structured hand gestures.

2 Related Work

Sign language recognition (SLR) can be categorized into three groups: (i) recognition of alphabets (fingerspelling) [15, 16], (ii) isolated words [5, 6, 7, 8] and (iii) continuous sentences [9, 10, 11]. In this paper, we focus on (ii), where each video represents one gesture.

Different methods to represent spatiotemporal features have been proposed in this field [5, 17, 6, 7, 8], including the use of two-stream CNNs [5, 17, 18]. In [6], the authors use 3D CNNs and LSTMs to encode global temporal and local spatial information. Afterwards, they capture global spatial information using 2D CNNs. Wang *et al.* [7] feed full-body and cropped-hand RGB and depth images. They fuse ConvNet based classification together with 3D ConvLSTMs based classification. Achieving the best results, Miao *et al.* [8] base their method on ResC3D leveraging both residual and C3D models. They propose a weighted frame unification strategy to sample key frames, and use a canonical correlation analysis-based fusion scheme for blending features.

3 Network Architecture

Our proposed pipeline for recognition of isolated sign language words is shown in Fig. 2, a two-stream model of inflated 3D (I3D) ConvNets. It is based on the architecture by [1]. The two-streams are fed RGB and optical flow data. We start by pre-trained weights on RGB and flow data from the Kinetics dataset [14]. Note that the two I3D networks do not share any parameters. We replace the final classification layer of each stream with outputs for the Chalearn IsoGD dataset, this layer's weights are randomly initialized.

I3D ConvNets were introduced by [1], they turn 2D ConvNet of the successful Inception-v1 architecture into a 3D convolutional counterpart. This is done by "inflating" 2D $k \times k$ kernels to 3D $t \times k \times k$ kernels that span over t frames. The kernels are initialized with pre-trained ImageNet weights [19], by creating a video that consists of a single static frame repeated over time in order to initialize 3D kernels from 2D ones, i.e. each of the t planes is initialized by the pre-trained $k \times k$ ImageNet weights and rescaled by $1/t$. The authors illustrate that I3D models outperform equivalent CNN+LSTM architectures.

We generate optical flow frames from the RGB data based on the "Dual TV-L1" optical flow algorithm in [20]. The flow data is used as another modality that represents the motion path in the videos. The intrinsic recurrence in optical flow algorithms allows it to better represent motion features in the videos, despite the use of 3D ConvNets, which perform mere feedforward computations [1].

Table 1. Recognition results on the validation subset of ChaLearn249 IsoGD for each of the RGB and optical flow streams, and a comparison with state-of-the-art methods using only RGB and/or optical flow modalities.

Modality	Method	Accuracy
RGB	XDETVP[6]	51.31%
	SYSU_ISEE[22]	47.29%
	ASU[8]	45.07%
	I3D-SLR (ours)	**54.63%**
Optical Flow	XDETVP[6]	45.30%
	ASU[8]	44.45%
	I3D-SLR (ours)	**54.84%**
RGB + Flow	SYSU_ISEE[22]	41.65%
	I3D-SLR (ours)	**62.09%**

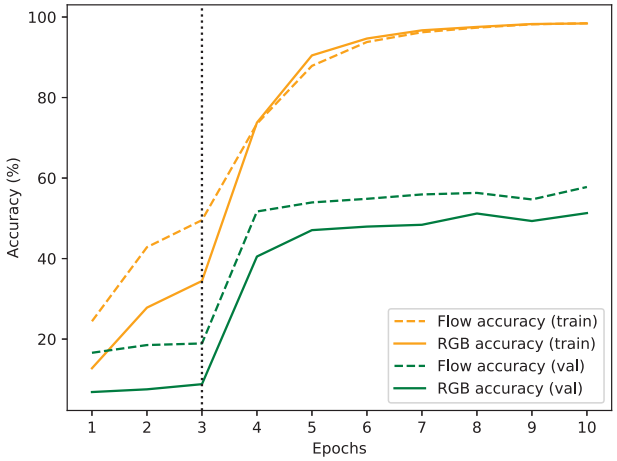

Fig. 3. The performance of different modalities at every epoch during training. The dotted line marks the point after which the random top layers have been trained, and fine-tuning of the entire network has started.

The generation of flow frames has been done as a pre-processing step prior to training. Afterwards, we uniformly downsample our videos to a fixed number of 40 frames per video since CNNs expect the number of frames for each video to be constant. We crop the frames around the center to a spatial size of 224×224.

Data augmentation is particularly important in our case since the dataset is signer-independent (see Sec. 4.1). Therefore, to increase the diversity of our training set, we perform spatial augmentation to our videos during training. This involves image shifts along both the x- and y-axes and changing brightness of the videos.

For training, we start by freezing the transferred weights and only training the randomly initialized top layers of each of the RGB and flow streams. The initial learning rate is set to 10^{-3} for 3 epochs. Afterwards, the learning rate is dropped by $1/10$ and the entire stream is fine-tuned. For stochastic optimization we opted for Adam [21] optimizer, and a mini-batch size of 4. We employ categorical cross-entropy as our loss function. Upon evaluation, the predictions from each stream are averaged together to output a single label for each video.

4 Experiments

4.1 Dataset and Metrics

We use ChaLearn249 IsoGD dataset [4] for our experiments. It is one of the latest, large-scale RGB-D SLR benchmarks recorded by a Microsoft Kinect camera. The total number of videos is 47,933 belonging to 249 classes performed by 21 different signers. Each video represents one gesture instance. One of the main challenges of the dataset is that it is signer-independent, i.e. the signers in each subset of the dataset are unique. However, this is fitting to our goal of designing a neural network that is applicable to other SLR tasks. We highlight that in this work we only use the RGB data of the dataset.

The dataset has a clear evaluation protocol and is already split into three mutually exclusive subsets: training (35,878 videos), validation (5,784 videos), and test (6,271 videos). In our experiments, we report and compare the accuracy on both validation and test sets as described for the ChaLearn249 IsoGD dataset as follows:

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n}\delta(p_l(i), t_l(i)), \tag{1}$$

where n is the number of samples, p_l is the predicted label, t_l is the ground truth, $\delta(j_1, j_2) = 1$ if $j_1 = j_2$, otherwise $\delta(j_1, j_2) = 0$.

4.2 Results

In this section we report results of evaluating the RGB and optical flow streams separately to highlight the significance of each stream, and together by concatenating predictions of each stream and predicting a certain label. Table 1 shows these results along with a comparison with recent state-of-the-art results on the ChaLearn249 dataset. We include approaches that only utilize RGB and/or optical flow data, without using depth data for a fair comparison. We outperform state-of-the-art methods by over 3% and 9.54% for the RGB and Optical flow modalities respectively. We only show validation results in Table 1 as the other methods do not provide results on the test set when using RGB only. For the test set, we report an accuracy of 57.73% for the RGB stream, 57.68% for the optical flow stream, and 64.44% when using both RGB and flow data.

We plot the performance of the two streams in terms of accuracy during training in Fig. 3. During the first 3 epochs of training, the weights pre-trained on the Kinetics dataset

Fig. 4. Left and Middle: Correctly-classified gestures. Right: Gesture 011 mostly misclassified as Gesture 026.

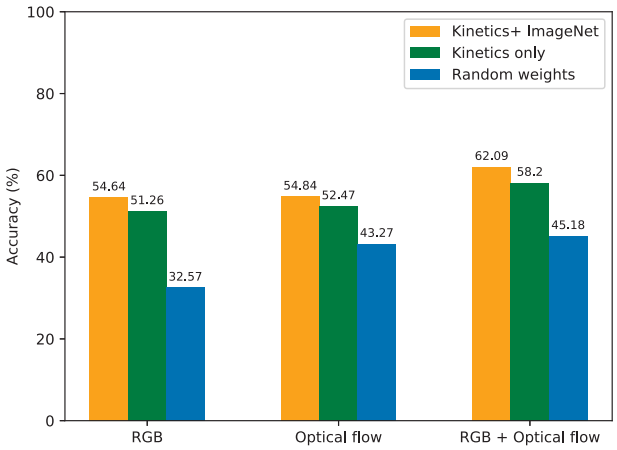

Fig. 5. Accuracy of different modalities with different weight initializations to assess the effect of transfer learning.

were frozen, and only the top layers with randomly initialized weights were being trained. Training was automatically stopped after 10 epochs by the early-stopping criterion.

In Fig. 4, we show samples of correctly-classified and misclassified gestures from our experiments. On the left, the same signers are performing the same gesture, Gesture 220, and are doing it differently; one signer is also using his thumb, while the other one is folding it away. This shows a typical case causing intra-class variations. Both inputs were correctly classified by our model as Gesture 220. Despite high inter-class similarity between gestures 001 and 221, Fig. 4 (middle), our proposed model was also successfully able to differentiate between the two. Gestures 011 and 026 are another example of high inter-class similarity, however, the model here fails to consistently differentiate between these two gestures, see Fig. 4 (right).

Since the aforementioned methods in Table 1 were not originally designed for use on RGB data only, rather relied on the use of depth modality as well, we briefly mention these results here. Miao *et al.* [8]. achieve state-of-the-art results reporting 64.40% and 67.71% validation and test accuracy respectively. *Without using depth data*, our proposed method achieves a close 62.09% for the validation set, and an accuracy of 64.44% for the test set. Our RGB-only results for the validation data rank 2nd in comparison to all state-of-the-art methods [8, 7, 6, 12] despite their usage of the depth modality as well. It should be noted that the baseline[1] [12] for the test set is very challenging, achieving an accuracy of 67.26%, which other state-of-the-art methods that utilize both RGB and depth modalities do not surpass [7, 6].

4.3 Different Weight Initializations

In this subsection we evaluate the effectiveness of transferring pre-trained action recognition weights to the task of SLR by conducting two experiments. First, we initialize our network with pre-trained weights using Kinetics only before fine-tuning to our sign language dataset. In other words, ImageNet weights were not used upon inflating the 2D kernels into 3D ones, instead they were randomly initialized and directly trained on Kinetics [1]. Second, we randomly initialize our network weights, and only train on ChaLearn IsoGD. For both experiments, we keep the hyperparameters setting as mentioned above [23]. Results for these experiments are shown in Fig. 5.

We observe that the results are only slightly lower, from an accuracy of 62.09% to 58.2% when using Kinetics dataset only. The decrease in results is more attributed to the lower performance of the RGB stream, more than 3% decrease in performance, than it is to the optical flow stream. This is explicable since ImageNet weights reflect no motion. However, there is a significant decline in performance, by almost 13%, when training our model from scratch on ChaLearn IsoGD. This clearly shows the impact of using pre-trained action recognition weights to our task, and how SLR can benefit from them.

5 Conclusion

This paper presents an effective method for recognition of isolated sign language words relying only on RGB data, which is particularly useful for applications where depth data is not available. We used two-stream inflated 3D ConvNets for RGB and optical flow data. We also highlight the significance of transferring learning for video data. We show that SLR benefits from the valuable spatiotemporal features that have been learned for the task of action recognition.

[1]The baseline results were provided for the ChaLearn 2017 Large Scale Isolated Gesture Recognition Challenge by Wan *et al.* [12]

6 References

[1] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the Kinetics dataset," in *CVPR*, 2017.

[2] O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep sign: enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs," *IJCV*, vol. 126, no. 12, pp. 1311–1325, 2018.

[3] N.A. Sarhan, Y. El-Sonbaty, and S. M Youssef, "HMM-based arabic sign language recognition using kinect," in *International Conference on Digital Information Management*, 2015.

[4] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, and S. Z. Li, "ChaLearn looking at people RGB-D isolated and continuous datasets for gesture recognition," in *CVPR ChaLearn Looking at People Workshop*, 2016.

[5] L. Pigou, S. Dieleman, P. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," in *ECCV ChaLearn Looing at People: Pose Recovery, Action/Interation, Gesture Recognition*, 2014.

[6] L. Zhang, G. Zhu, P. Shen, J. Song, S. Afaq Shah, and M. Bennamoun, "Learning spatiotemporal features using 3DCNN and convolutional LSTM for gesture recognition," in *ICCV*, 2017.

[7] H. Wang, P. Wang, Z. Song, and W. Li, "Large-scale multimodal gesture recognition using heterogeneous networks," in *ICCV*, 2017.

[8] Q. Miao, Y. Li, W. Ouyang, Z. Ma, X. Xu, W. Shi, and X. Cao, "Multimodal gesture recognition based on the ResC3D network," in *ICCV*, 2017.

[9] O. Koller, S. Zargaran, and H. Ney, "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs," in *CVPR*, 2017.

[10] N. Cihan Camgoz, S. Hadfield, O. Koller, and R. Bowden, "SubUNets: End-to-end hand shape and continuous sign language recognition," in *ICCV*, 2017.

[11] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in *CVPR*, 2017.

[12] J. Duan, J. Wan, S. Zhou, X. Guo, and S. Z. Li, "A unified framework for multi-modal isolated gesture recognition," *ACM Trans. on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 1s, pp. 21, 2018.

[13] L. Pigou, A. Van Den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre, "Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video," *IJCV*, vol. 126, no. 2-4, pp. 430–439, 2018.

[14] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., "The Kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[15] C. Dong, M.C . Leu, and Z. Yin, "American sign language alphabet recognition using Microsoft Kinect," in *CVPR Workshops*, 2015.

[16] M. Mohandes, S. Aliyu, and M. Deriche, "Arabic sign language recognition using the leap motion controller," in *2014 IEEE 23rd Int'l Symposium on Industrial Electronics (ISIE)*. IEEE, 2014.

[17] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," in *ECCV ChaLearn Looing at People: Pose Recovery, Action/Interaction, Gesture Recognition*, 2014.

[18] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[20] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-L1 optical flow," in *Proceedings of the 29th DAGM Conference on Pattern Recognition*. Springer, 2007.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] Benchao Li, Wanhua Li, Yongyi Tang, Jian-Fang Hu, and Wei-Shi Zheng, "GL-PAM RGB-D gesture recognition," in *2018 25th IEEE ICIP*. IEEE, 2018, pp. 3109–3113.

[23] K. He, R. Girshick, and P. Dollar, "Rethinking ImageNet pre-training," in *ICCV*, October 2019.