Using Convolutional 3D Neural Networks for User-Independent Continuous Gesture Recognition

Necati Cihan Camgoz, Simon Hadfield
University of Surrey,
Guildford, GU2 7XH, UK
{n.camgoz, s.hadfield}@surrey.ac.uk

Oscar Koller
Human Technology & Pattern Recognition
RWTH Aachen University, Germany
koller@cs.rwth-aachen.de

Richard Bowden
University of Surrey,
Guildford, GU2 7XH, UK
r.bowden@surrey.ac.uk

***Abstract*—In this paper, we propose using 3D Convolutional Neural Networks for large scale user-independent continuous gesture recognition. We have trained an end-to-end deep network for continuous gesture recognition (jointly learning both the feature representation and the classifier). The network performs three-dimensional (i.e. space-time) convolutions to extract features related to both the appearance and motion from volumes of color frames. Space-time invariance of the extracted features is encoded via pooling layers. The earlier stages of the network are partially initialized using the work of Tran et al. before being adapted to the task of gesture recognition. An earlier version of the proposed method, which was trained for 11,250 iterations, was submitted to ChaLearn 2016 Continuous Gesture Recognition Challenge and ranked 2nd with the Mean Jaccard Index Score of 0.269235. When the proposed method was further trained for 28,750 iterations, it achieved state-of-the-art performance on the same dataset, yielding a 0.314779 Mean Jaccard Index Score.**

I. Introduction

Gestures are a natural form of human communication. As the use of computers has become more ubiquitous in our daily lives, human-computer interfaces have started to mimic natural human communication, allowing users to employ gestures to convey their intentions to computers. Hand and arm gestures are now widely used cues for human-computer interactions [1]. Although gestures are more natural, the use of gestures comes with its drawbacks, as imperfect human pose detection coupled with inter and intra-user spatio-temporal variability of gestures make it difficult to perform user independent gesture recognition.

Gesture recognition systems aim to detect and recognize gestures from a limited gesture vocabulary given a sensory input [2]. In early work, color cameras were widely used for the development of gesture recognition systems [3]. However, human pose estimation from color images is susceptible to color ambiguity between the user and the background, and this led researchers to use colored gloves [4]. With the emergence of consumer depth cameras [5], researchers quickly incorporated depth sensors into their systems, as depth simplifies the task of human pose estimation [6]. Many state-of-the-art gesture recognition systems today use depth images as a modality or as a means of preprocessing their data before recognizing gestures [2], [7], [8].

The use of data gloves was also proposed for gesture recognition. However, due to their high cost and calibration requirements, they did not become as popular as video-based gesture recognition systems [1].

Video-based gesture recognition typically starts with the acquisition and then extraction of meaningful features. Although the type of features may vary for each task, most of the systems use features that describe the users' upper body pose, hand shape and hand movements. These features are then used in conjunction with statistical learning methods to distinguish classes of gestures from each other.

Classification of gestures relies heavily on learning temporal (i.e. trajectory, speed) and spatial (i.e. visual features such as hand shape and hand location) aspects of gesture samples. Due to the spatio-temporal nature of gestures, modeling the temporal aspect plays a crucial role in gesture recognition. In the literature, there are several common approaches used to represent the temporal aspects of gestures:

The first approach is to discard the temporal order of a gesture and represent it by distributions of spatial features. Hernandez-Vela et al. proposed an adaptation of Bag of Words methods to recognize hand gestures, termed a Bag-of-Visual-and-Depth-Words (BoVDW) [9]. In a recent Chinese Sign Language Recognition study [10], Wang et al. proposed averaging the spatial feature covariance matrices extracted from each frame of the gesture sequences. Covariance matrices are then used to calculate distances between gesture samples in the Grassmannian Manifold. Although these methods may be suitable for small recognition tasks, they are unable to distinguish among similar gestures with different temporal ordering.

One of the most common approaches to modeling the temporal aspect of gestures is to represent gestures with spatio-temporal grammars. In these approaches, spatial features are grouped into the building blocks of gestures, such as states [11] or subunits [12], and changes among these states are modeled using graphical models. Since the pioneering work of Starner and Pentland [3], Hidden Markov Models have often been used for gesture recognition [13], [14], [15]. Other graphical models such as Hidden Conditional Random Fields [16], Autoregressive Models [17] and Recurrent Neural Networks [8], [18] have also been deployed for the gesture recognition task.

Another approach for temporal modeling is using gesture templates. Instead of grouping spatial features into clusters and learning the interactions between these clusters, these models learn static sequential patterns of features called templates [19].

Templates are often constructed by stacking or concatenating a fixed number of spatial features over the temporal domain. Motion History Images are an example of these approaches [20]. As templates have no mechanism to represent the execution speed of a gesture, templates should either be resampled to represent different execution speeds, or they should be temporally aligned using methods such as Dynamic Time Warping [21].

With the availability of large annotated datasets, deep learning methods have become a feasible solution for gesture recognition. In recent years, Convolutional Neural Network (CNN) based approaches have achieved state-of-the-art performance in gesture recognition challenges [7], [8]. In [7], Neverova et al. proposed a multi-scale and multi-modal deep learning architecture to spot and recognize continuous gestures, and achieved state-of-the-art performance in the ChaLearn 2014 Gesture Recognition challenge [22]. In [8], Pigou et al. proposed temporally modeling the spatial features obtained from CNNs by using Recurrent Neural Networks (RNNs) with Long Short-Term Memory units, and shows the benefits of using RNNs over temporal pooling approaches. In a recent study, 3D Convolutional Neural Networks were proposed for the isolated hand gesture recognition task using depth and intensity modalities in automotive interfaces and reported a better recognition performance than using HOG descriptors [23].

In this paper, we propose using 3D Convolutional Neural Networks (3D CNNs), which are capable of learning both the spatial and the temporal aspects of the data, for user-independent large scale continuous gesture recognition tasks. We train an end-to-end deep neural network for both feature learning and classification from color video sequences. Each frame of a gesture sequence is represented by a spatio-temporal volume of its neighboring 16 color frames. We apply a sliding volume approach over the color videos and obtain class probabilities of each frame. These probabilities are then subjected to two layers of majority filtering before assigning the final label of each frame. We partially initialize our networks with the work of Tran et al. [24].

To find the best performing parameters, we have performed various experiments to evaluate the effect of mirroring the training data, applying models that were pre-trained on different datasets and the number and initialization of layers that will be fully trained instead of being fine-tuned.

The rest of the paper is structured as following: In Section II, we examine the deep learning applications in the field of gesture recognition and describe the proposed method. In Section III, we give details of the ConGD, which was introduced by Wan et al. [25] for the ChaLearn 2016 Large Scale Continuous Gesture Recognition Challenge. In Section IV, we share our experimental setup and results. Finally, we conclude our paper in Section V by discussing our findings and future studies.

II. 3D Convolutional Neural Networks

Gestures can be defined as a time-based sequence of spatial configurations and disregarding either the spatial or the temporal information can result in poor performance in recognition.

In the past, hand-crafted features have been used to describe the spatial information of the gestures. These hand-crafted features include, but are not limited to, hand shape descriptors and upper body pose information. These descriptors are then tracked through the time domain using approaches such as graphical models to represent the temporal aspect of the gesture.

Inspired by the recent progress in the field of deep learning 2D Convolutional Neural Networks (2D CNNs) have been applied to the Gesture Recognition field in order to extract spatial features [7], [8]. In recent studies, features were either concatenated into fixed sized gesture templates [7] or passed to HMM [13] or Recurrent Neural Networks [8] in order to model the temporal aspects of the gestures.

In a more recent study, 3D Convolutional Neural Networks (3D CNNs) were proposed to recognize isolated gestures [23]. Depth and intensity information were combined into a single image and these in turn combined to form gesture volumes. The gesture volumes are then resampled to have fixed size before being used for training 3D CNNs. In comparison to 2D CNNs that are capable of learning the spatial information from single images, 3D CNNs can learn both the spatial and the temporal information from a sequence of images, thus eliminating the need for secondary temporal modeling techniques.

In this paper, we propose using 3D CNNs for user-independent continuous gesture recognition. The initial architecture is based on that proposed by Tran et al. for action recognition [24]. The network architecture consists of 8 3D convolutional layers, five 3D max-pooling layers, two fully connected layers and a softmax classification layer (See Figure 1).

To be able to do frame-wise classification, we have constructed volumes for a given frame (F_t) by using its surrounding 16 frames ($F_{t-7:t+8}$). Volumes were created for each gesture sequence in a sliding manner, with each volume having the label of the gesture occurring at its central frame.

More formally, the first layer of the spatio-temporal CNN is defined as

$$\text{Conv1}(x,y,t) = \sum_{\delta x,\delta y,\delta t} F_{t+\delta t}(x+\delta x, y+\delta y) \times w(\delta x, \delta y, \delta t)\,, \tag{1}$$

where x and y define the pixel position within frame t. The spatio-temporal neighbourhood that δx, δy and δt are drawn from is defined by the kernel size of the convolutional layers ($3 \times 3 \times 3$ in our experiments). Note that the output domain of the convolution is still three dimensional, maintaining the spatio-temporal arrangement of the learned patterns.

The common "Rectified linear unit" approach is used to inject nonlinearities into the learning process. Thus

$$\text{ReLU1}(x,y,t) = \begin{cases} \text{Conv1}(x,y,t) & \text{if Conv1}(x,y,t) > 0 \\ 0 & otherwise\,. \end{cases} \tag{2}$$

Finally, the first spatio-temporal pooling layer is then defined as

$$\text{Pool1}(x,y,t) = \max_{x,y,t} (\text{ReLU } 1(x,y,t))\,, \tag{3}$$

where in this case the x, y and z neighbourhoods relate to the scale of the pooling ($2 \times 2 \times 2$ for all but the first layer in our

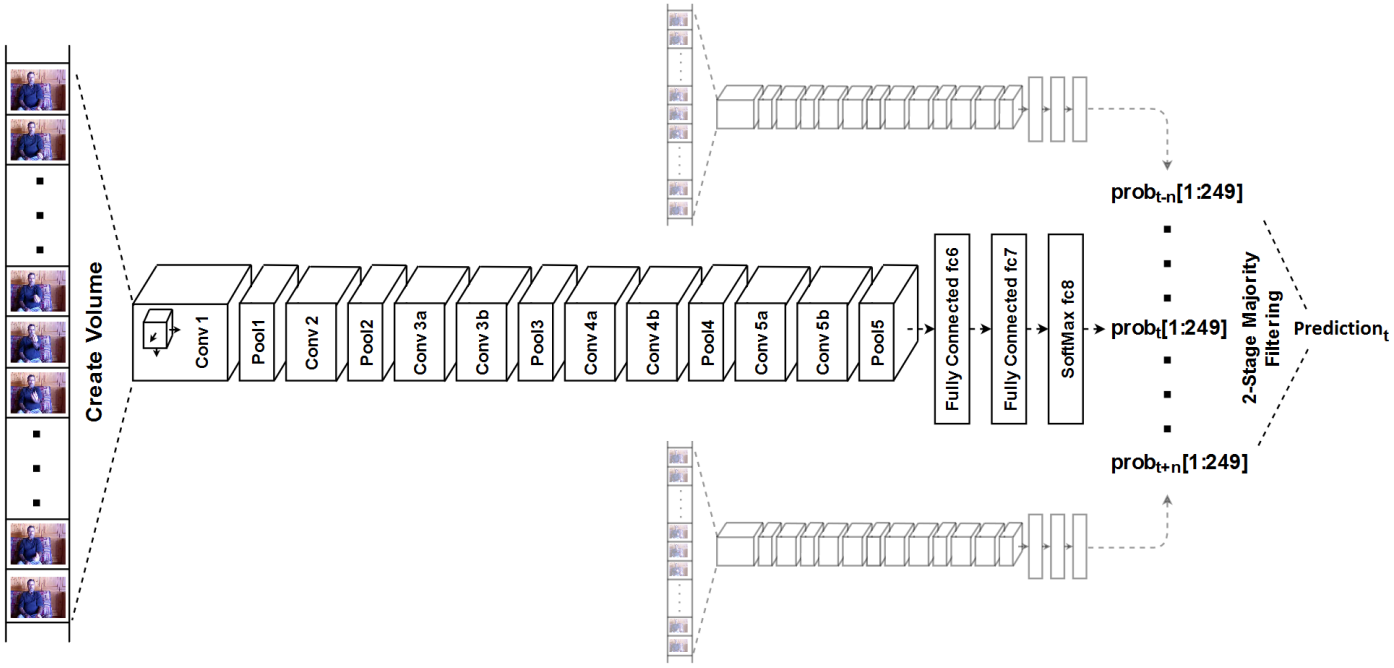

Fig. 1. Overview of the proposed framework.

experiments). This spatio-temporal helps to provide robustness to temporal variations; distinctive patterns are encoded in the same manner, regardless of where they occur within a local region of the spatio-temporal volume.

In order to remove the noise from the final predictions of the network, we have applied 2-stage majority filtering. Using the class probabilities extracted from the deep network, we first apply a majority filter with the size of 33 frames to the prediction of gesture sequence frames using a sliding window. Finally, we apply a second majority filter with a size of 17 frames to the output of the first majority filter, thus removing the noise in a coarse to fine manner. The majority filter sizes were chosen empirically to maximize the frame-based accuracy.

III. ChaLearn 2016 Continuous Gesture Dataset

The Continuous Gesture Dataset (ConGD) [25], featured by ChaLearn 2016, was designed to evaluate the performance of user-independent gesture recognition methods. The dataset was originally collected for the ChaLearn 2011 Gesture Recognition Challenge [26], but the new protocol was introduced by Wan et al. to allow researchers to evaluate their methods for user-independent recognition. ConGD consists of 47,933 gesture samples belonging to 249 gesture classes, making it the largest user-independent dataset [27] surpassing the ChaLearn 2014 Gesture Recognition Dataset [22] which has 13,858 samples and the DeviSign Chinese Sign Language Dataset which has 24,000 samples [28].

ConGD has 21 subjects that are separated into Train, Validation and Test partitions in a mutually exclusive manner. A summary of the data partitions can be seen in Table I.

ConGD was recorded by Microsoft Kinect sensor [5]. It only includes color and Depth video sequences provided by the sensor, making it more challenging from the other datasets [22], [28] collected by Kinect as it does not provide the human pose information.

TABLE I
ChaLearn 2016 ConGD Partition Information

Partition Name	# of Samples	# of Sequences	# of Subjects
Train	30,442	14,134	17
Validation	8,889	4,179	2
Test	8,602	4,042	2
All	47,933	22,355	23

ConGD was proposed by Wan et al. in [25] for the ChaLearn 2016 Continuous Gesture Recognition Challenge and a baseline was presented for the dataset which uses Mixed Features around Sparse Keypoints (MFSK) and Bag of Visual Words based approach (See Table II).

IV. Experiments & Results

A. Model Selection

In order to find the best performing setup, experiments were performed to examine the effects of mirroring input volumes, initializing fully connected layers instead of using the pre-trained layers, and initializing the softmax layer with different distributions. While searching for the best performing setup, we have used the model which was trained on Sports1m [29] for 1.9 million iterations by Tran et al. [24] as our basis model.

The generalization capability of deep learning methods heavily relies on the data it has been trained on. To make the best use of available datasets, it is common to include ”augmentations” which inject additional variance into the data without requiring additional collection or labeling time. For some tasks, the handedness of the user is not important. For gestures, users

might choose their left or right hand while performing a gesture, without changing the meaning. Therefore both left, and right-handed gestures should be taken into consideration while training the system. One solution is to sometimes vertically mirror the training data, so the system is exposed to a wider variety of left and right-handed gestures. To evaluate the effects of mirroring the data, two models were trained. The first with no mirroring and a second model where the mirrored samples were added to the training.

In Figure 2 we look at the effect of mirroring the data on the frame based accuracy as the training proceeds. As can be seen, mirroring the training data provides a significant improvement in recognition performance across all iterations as expected for gesture recognition.

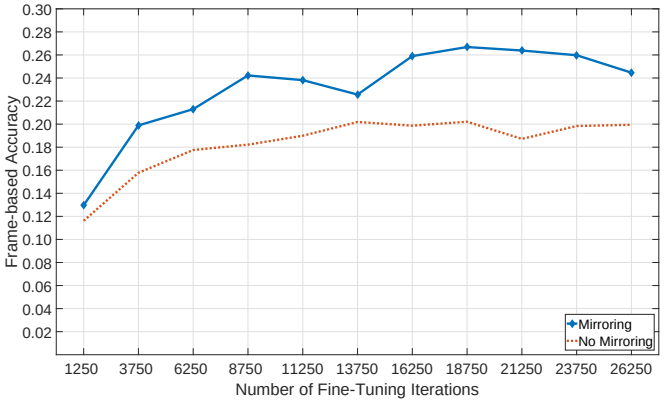

Fig. 2. The effect of mirror augmentation during training.

Initialization of layers has an immense effect on the optimization of the weights in a deep network. Proposed by Grolot et al. [30], *Xavier* initialization help weak signals to reach deeper in the network. Instead of the using Gaussian Distribution (0 Mean, 0.005 Standard Deviation) as in previous related work on action recognition, we have initialized the softmax layer using *Xavier* initialization and examined its effects.

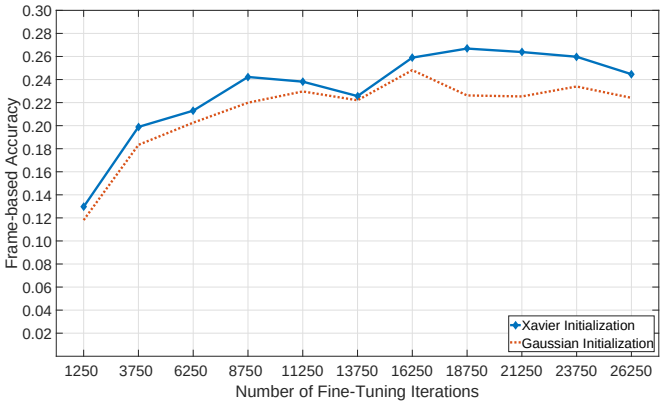

Fig. 3. The effect of intelligent weight initialization.

As it can be seen in Figure 3, initializing the softmax layer with *Xavier* initialization improves the frame-based accuracy throughout all of the iterations. Therefore, in our remaining experiments we have used *Xavier* rather than Gaussian initialization.

The later fully connected layers tend to be more specialized to the task which the networks was trained on. It is reasonable to question whether these layers have already converged to a minima which is unsuitable for a new task, in which case better performance might be obtained by initializing and training these layers from scratch.

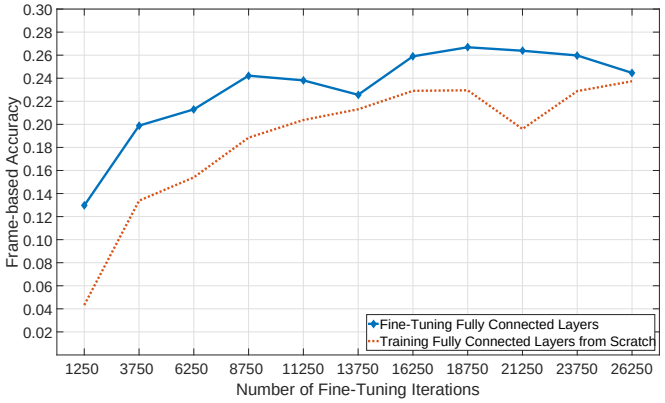

Fig. 4. Pre-training of classification layers on a different task.

Counter to this reasoning, Figure 4 shows that fine-tuning the pre-trained layers instead of initializing them performed drastically better across all iterations. Based on these results we have not initialized the fully connected layer in the following experiments.

One of the cornerstones of deep learning is the ability to exploit vast quantities of training data. As a result, it is often beneficial to pre-train networks, even on tasks which are seemingly very different to the intended task. The action recognition network of Tran et al. was initially trained on the Sports1m dataset [29]. We compare this to a second network, which has been exposed to both the Sports1m and UCF101 [31] action recognition datasets.

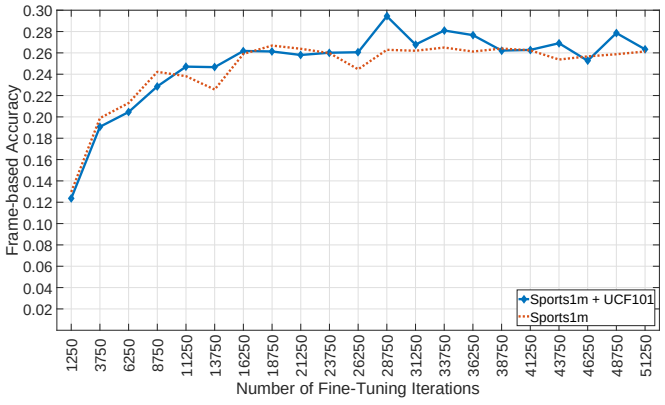

Fig. 5. Pre-training on other datasets.

As shown in Figure 5, the additional exposure to a wider range of actions leads to some improvements in our continuous

gesture recognition task. This is despite the fact that there is little overlap between the classes of the two action recognition datasets, and there is no overlap with the ChaLearn dataset. It appears that the low-level spatio-temporal patterns learned by the network become increasingly generic as they are exposed to more data from different tasks.

B. Model Training

In light of our previous experiments, we fine-tuned our model from the weights that were pre-trained on Sports1m and UCF101. We added mirrored samples to the training data and initialized the softmax layer using *Xavier* initialization. The model was fine-tuned for *100,000* iterations with a learning rate of *0.001* and *0.9* momentum.

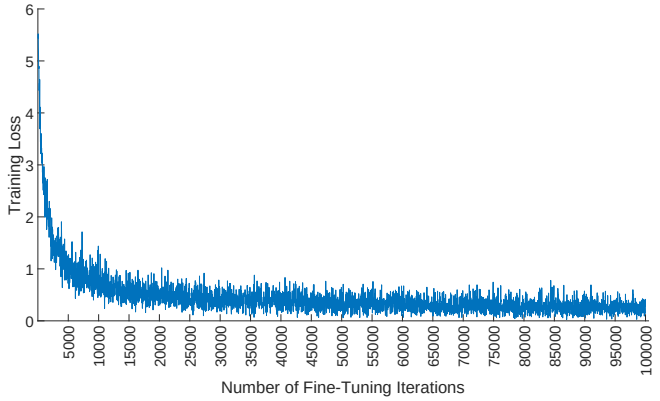

Fig. 6. Training loss over 100,000 iterations.

As it can be seen from Figure 6 the training started to converge after 15,000 iterations and converged after 60,000 iterations. We evaluated the performance of the system every 1,250 iterations on both the training and validation sets of ConGD and chose the best performing method depending on the frame based accuracy they yielded.

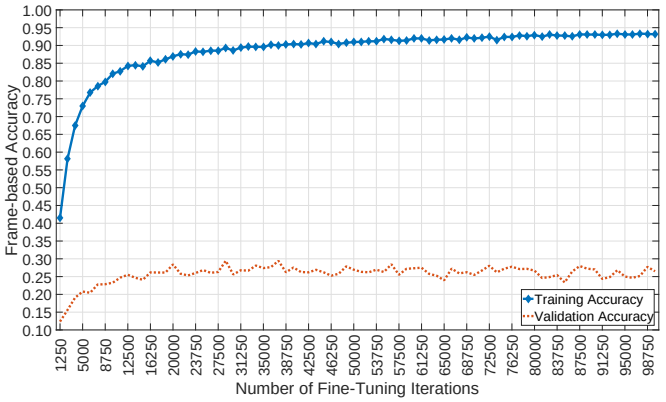

Fig. 7. Train and validation accuracy over 100,000 iterations.

Figure 7 shows us the training and validation set accuracies across all the iterations. The validation accuracy had a fast rising trend until it achieved a peak score of 0.2934 frame-based accuracy, followed by a stabilized 30,000 iterations and finally started a slow decreasing trend. However, the training accuracy kept rising indicating an overfit to the training samples. We have chosen the best performing method that was trained for 28,750 iterations and compared it with the participants of the Chalearn 2016 Continuous Gesture Recognition challenge.

C. Comparing the Performance with Challenge Participants

Due to time constraints the version that was submitted to the ChaLearn 2016 gesture recognition challenge was ranked second, achieving a mean Jaccard Index score of 0.269 on the test data (See Table II). Recognition performance of the top 3 competitors was announced by the challenge organizers and all of the methods used a variety of deep neural networks. The first ranking method, proposed by Chai et al. [18], is based on two stream recurrent neural networks which use multimodal features, and achieved 0.287 mean Jaccard Index score. The method proposed by Wang et al. [32], which ranked third, proposed extracting improved depth motion maps from depth sequences and classifying them using 2D CNNs. When we subsequently fully trained our method (with an additional day of training) it surpassed the first ranking method by achieving a mean Jaccard Index score of 0.315.

TABLE II
CHALEARN 2016 CONGD CHALLENGE RESULTS (MJI: MEAN JACCARD INDEX)

Rank	Team	Method	Validation MJI	Test MJI
N/A	Baseline [25]	MFSK	0.090200	0.146400
3	AMRL [32]	IDMM + CNN	N/A	0.265506
2	TARDIS (Ours)	3D CNN	0.280860	0.269235
1	ICT_NHCI [18]	RNNs	N/A	0.286915
N/A	TARDIS (Best)	3D CNN	0.342971	0.314779

V. CONCLUSION

In this study, we have proposed applying 3D Convolutional Neural Networks (3D CNNs) to the problem of large-scale continuous user-independent gesture recognition. Compared to previous deep architectures what were proposed for this task, 3D CNNs are capable of encoding the spatial and temporal information in the data without requiring additional temporal modeling. The 3D convolution and pooling layers help to learn the spatio-temporal variations in the data.

Our experiments have shown that mirroring improves the gesture recognition performance drastically in tasks where there is no dominant hand (i.e. the same gestures can be performed using either hand). In light of our experiments, it is obvious that the quality of the spatio-temporal patterns (and thus the performance) provided by 3D CNNs, is improved with exposure to more training data. This is even true for the higher fully connected "classification" layers, and even when the tasks being trained for are different. If layers must be initialized from scratch, careful choice of weight initialization can also significantly improve performance.

We have applied the proposed method to the ConGD dataset, which was introduced with the ChaLearn 2016 Continuous Gesture Recognition Challenge, and ranked 2nd with a model which was fine-tuned for 11,250 iterations. After the end of

the challenge we have obtained state-of-the-art recognition performance on the same dataset with a model which had the same parameters but was trained for 28,750 iterations.

As future work, it would be interesting to further investigate the effects of data exposure, not only from different tasks but even from different data modalities.

ACKNOWLEDGMENT

This work was funded by the SNSF Sinergia project "Scalable Multimodal Sign Language Technology for Sign Language Learning and Assessment (SMILE)" grant agreement number CRSII2_160811.

REFERENCES

[1] S. Mitra and T. Acharya, "Gesture Recognition: A Survey," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 37, no. 3, pp. 311–324, 2007.

[2] S. S. Rautaray and A. Agrawal, "Vision-based Hand Gesture Recognition for Human Computer Interaction: a Survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1–54, 2015.

[3] T. Starner and A. Pentland, "Real-time American Sign Language Recognition from Video using Hidden Markov Models," *Motion-Based Recognition*, pp. 227–243, 1997.

[4] O. Aran, "Vision-based Sign Language Recognition: Modeling and Recognizing Isolated Signs with Manual and Non-Manual Components," Ph.D. dissertation, Bogazici University, 2008.

[5] Z. Zhang, "Microsoft Kinect Sensor and Its Effect," 2012.

[6] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR*, 2011.

[7] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "ModDrop: Adaptive Multi-modal Gesture Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014.

[8] 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," *arXiv preprint arXiv:1506.01911*, 2015.

[9] A. Hernández-Vela, M. A. Bautista, X. Perez-Sala, V. Ponce, X. Baró, O. Pujol, C. Angulo, and S. Escalera, "BoVDW: Bag-of-Visual-and-Depth-Words for gesture recognition," in *IEEE International Conference on Pattern Recognition (ICPR)*, 2012.

[10] H. Wang, X. Chai, X. Hong, G. Zhao, and X. Chen, "Isolated Sign Language Recognition with Grassmann Covariance Matrices," *ACM Transactions on Accessible Computing*, vol. 8, no. 4, 2016.

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

[12] H. Cooper, E.-J. Ong, N. Pugeault, and R. Bowden, "Sign Language Recognition using Sub-Units," *Journal of Machine Learning Research*, vol. 13, pp. 2205–2231, Jul 2012.

[13] O. Koller, H. Ney, and R. Bowden, "Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] R. Yang and S. Sarkar, "Gesture Recognition using Hidden Markov Models from Fragmented Observations," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 766–773.

[15] C. Keskin, A. Erkan, and L. Akarun, "Real-time Hand Tracking and 3D Gesture Recognition for Interactive Interfaces using HMM," *ICANN/ICONIPP*, pp. 26–29, 2003.

[16] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, "Hidden Conditional Random Fields for Gesture Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 1521–1527.

[17] T. Ishihara and N. Otsu, "Gesture Recognition using Autoregressive Coefficients of Higher-order Local Auto-correlation Features," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2004, pp. 583–588.

[18] X. Chai, Z. Liu, F. Yin, Z. Liu, and X. Chen, "Two Streams Recurrent Neural Networks for Large-Scale Continuous Gesture Recognition," in *International Conference on Pattern Recognition Workshops*, 2016.

[19] N. C. Camgöz, A. A. Kindiroglu, and L. Akarun, "Gesture Recognition using Template based Random Forest Classifiers," in *Workshop at the European Conference on Computer Vision*, 2014, pp. 579–594.

[20] J. W. Davis, "Hierarchical Motion History Images for Recognizing Human Motion," in *IEEE Workshop on Detection and Recognition of Events in Video*, 2001, pp. 39–46.

[21] X. Chai, G. Li, Y. Lin, Z. Xu, Y. Tang, X. Chen, and M. Zhou, "Sign Language Recognition and Translation with Kinect," in *International Conference on Automatic Face and Gesture Recognition*, 2013.

[22] S. Escalera, X. Baró, J. Gonzalez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H. J. Escalante, J. Shotton, and I. Guyon, "Chalearn Looking at People Challenge 2014: Dataset and Results," in *Workshop at the European Conference on Computer Vision*. Springer, 2014, pp. 459–473.

[23] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 1–7.

[24] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[25] J. Wan, S. Z. Li, Y. Zhao, S. Zhou, I. Guyon, and S. Escalera, "ChaLearn Looking at People RGB-D Isolated and Continuous Datasets for Gesture Recognition," in *Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[26] I. Guyon, V. Athitsos, P. Jangyodsuk, H. J. Escalante, and B. Hamner, "Results and Analysis of the ChaLearn Gesture Challenge 2012," in *Advances in Depth Image Analysis and Applications*. Springer Berlin Heidelberg, 2013, pp. 186–204.

[27] S. Ruffieux, D. Lalanne, E. Mugellini, and O. A. Khaled, "A Survey of Datasets for Human Gesture Recognition," in *International Conference on Human-Computer Interaction*. Springer, 2014, pp. 337–348.

[28] X. Chai, H. Wanga, M. Zhoub, G. Wub, H. Lic, and X. Chena, "DEVISIGN: Dataset and Evaluation for 3D Sign Language Recognition," Beijing, Tech. Rep., 2015.

[29] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[30] X. Glorot and Y. Bengio, "Understanding the Difficulty of Training Deep Feedforward Neural Networks," in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[31] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild," *arXiv preprint arXiv:1212.0402*, 2012.

[32] P. Wang, W. Li, S. Liu, Y. Zhang, Z. Gao, and P. Ogunbona, "Large-scale Continuous Gesture Recognition Using Convolutional Neutral Networks," *International Conference on Pattern Recognition Workshops*, 2016.