SHREC 2017 : Hand gesture classification challenge. Classify sequence by key frames with convolutional neural network

Joris Guerry* Bertrand Le Saux* David Filliat†

***Abstract*— The 3D Hand Gesture Recognition track of the 2017 SHREC challenge consists of classifying the gesture performed in sequences of depth images that can vary from 15 to 150 images. Gestures are categorized into 14 classes, 28 if one takes into account the fact that the hand can be opened or closed. Although the hand skeleton information is available in the dataset we opted for a solution based solely on the depth map. We obtain an accuracy of 71.9% on the 28 classes classification task, which goes up to 82.9% if we consider only the 14 classes.**

I. Encoding

Our approach is based on Convolutional Neural Networks (CNNs). This is why we sought an encoding that could adapt to the varying length of the sequences and that could be stored as a tensor. Intuitively, one can note it is possible to guess the gesture by just looking at subsets of images of the sequence. Based on this observation we thought to simply concatenate depth images on each other: each channel representing a keyframe. An exemple is shown for 3-keyframe concatenation in Fig.[I].

Fig. 1. 3 keyframe concatenation of a grab gesture

II. Method

The tensor previously defined is then processed by a CNN. Namely, we took a VGG11 network [1] whose parameters were initiated with weights resulting from ImageNet [2] training.

1) We first learnt to classify the 14 classes (giving much better results than starting by 28 classes classification)
2) Then we use the same features until "conv5" layer to learn a second binary task : is the hand open or not ? (Which is directly related to the 28 classes problem : $class_{28} = 2 * class_{14} - \delta$, where δ corresponds to 1 if the hand is closed and 0 otherwise.)
3) We could then use the 14 classes predictions ("fc8_14" layer) and the binary predictions ("fc8_δ" layer) to predict the 28 classes with a last fully connected layer.
4) Lastly, we get better results by removing the multi-task training (14 and δ then 28) and just training on the 28 classes prediction.

*ONERA The French Aerospace Lab, DTIM, F-91761 Palaiseau, France surname.name@onera.fr
†ENSTA Paristech, U2IS, F-91762 Palaiseau, France surname.name@ensta-paristech.fr

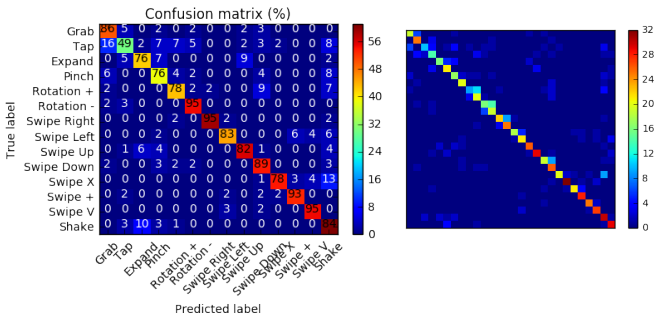
III. Results

The keyframes are picked regularly with a random variation up to 3 neighboring frames. We tried 3, 5 and 10 keyframes, 5 showing the best performance.

TABLE I
Accuracy on test dataset

Task	Accuracy	Mean computation time
28 classes	71.9%	236ms
14 classes (from 28 classes)	82.9%	236ms
14 classes (direct prediction)	81.0%	81ms

Fig. 2. Confusion matrix on 14 classes task (left) and 28 classes (right) :



(a) Confusion matrix (14 classes) (b) Confusion matrix (28 c.)

IV. Conclusion

We proposed a simple method to classify gesture sequences, using only 5 concatenated depth images as input of a neural network which yields to 82.9% accuracy on the 14 classes task. Improvment could be done by data augmentation strategies (mixing sequences) and by averaging results of different keyframes selections for a same sequence.

References

[1] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009.