EUROGRAPHICS Workshop on ... (2016), pp. 1–1
J. Jorge and M. Lin (Guest Editors)

3D Model Signatures via Aggregation of Multi-view CNN Features

Yangyan Li[1,2], Noa Fish[1], Daniel Cohen-Or[1] and Baoquan Chen[2]
[1]Tel Aviv University, Israel [2]Shandong University, China

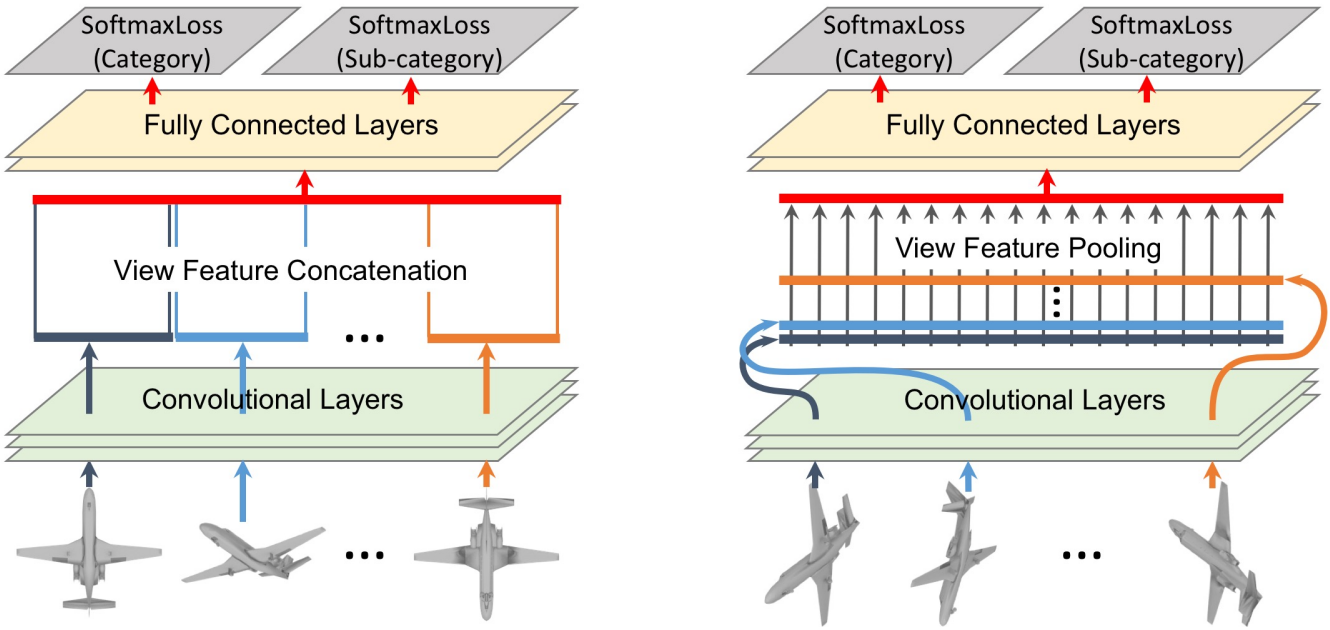

Figure 1: *CNN architectures for extracting 3D model signatures by aggregating CNN features of the multi-view images rendered from the 3D models. Mutli-view CNN features are aggregated by concatenating the CNN extracted view features when consistent 3D model orientation is avaiable (left), or max-pooling them when the model orientations are unknown (right).*

1. Method Description

3D models can be rendered into multiple view images, thus their signatures can be represented by the multi-view image features [CTSO03]. In this way, the informativeness and discriminativeness of the image features play critical role in the quality of the 3D model signatures. In recent few years, CNN extracted image features have been shown to be very successful in image based recognition tasks, as they are both informative and discriminative. Such advances in image feature extraction can be levereged to boost the performance of 3D model signatures.

Following [SMKLM15], we represent 3D model signatures with CNN features from multi-view images rendered from the 3D models. We first extract CNN features for each rendered view with the convolutional layers of a CNN model fine tuned for classifying individual rendered view into the category and sub-category of the 3D models, then aggregate the view features, and finally train several fully connected layers on classification tasks based on the aggregated features.

The key of this approach is the aggregation of CNN features from different views. In [SMKLM15], the view features are aggregated by a max-pooling, which treats different views evenly, thus is independent of how the rendering views are chosen. This is an advantage when the model orientations are unknown, but a disadvantage when the models are already consistently aligned. An aggregation method that is aware of the consistency should perform better on consistently aligned models. Instead of using max-pooling, we aggregate multi-view CNN features by concatenating the view features for 3D models with consistent orientations. In the concatenation based view feature aggregation, the fully connected layers can leverage the consistency of the rendered views, and make decisions based on an holistic understanding of all views. See Figure 1 for these CNN architectures.

More specifically, we render each model from 12 evenly distributed views, and fine tune VGG-19 network [SZ14] for the convolutional layers that is used for view feature extraction. We open source our code at `http://yangyanli.github.io/SHREC2016/` for reproducing our results.

References

[CTSO03] Chen D.-Y., Tian X.-P., Shen Y.-T., Ouhyoung M.: On visual similarity based 3d model retrieval. In *Computer graphics forum* (2003), vol. 22, Wiley Online Library, pp. 223–232. 1

[SMKLM15] Su H., Maji S., Kalogerakis E., Learned-Miller E.: Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 945–953. 1

[SZ14] Simonyan K., Zisserman A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014). 1