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

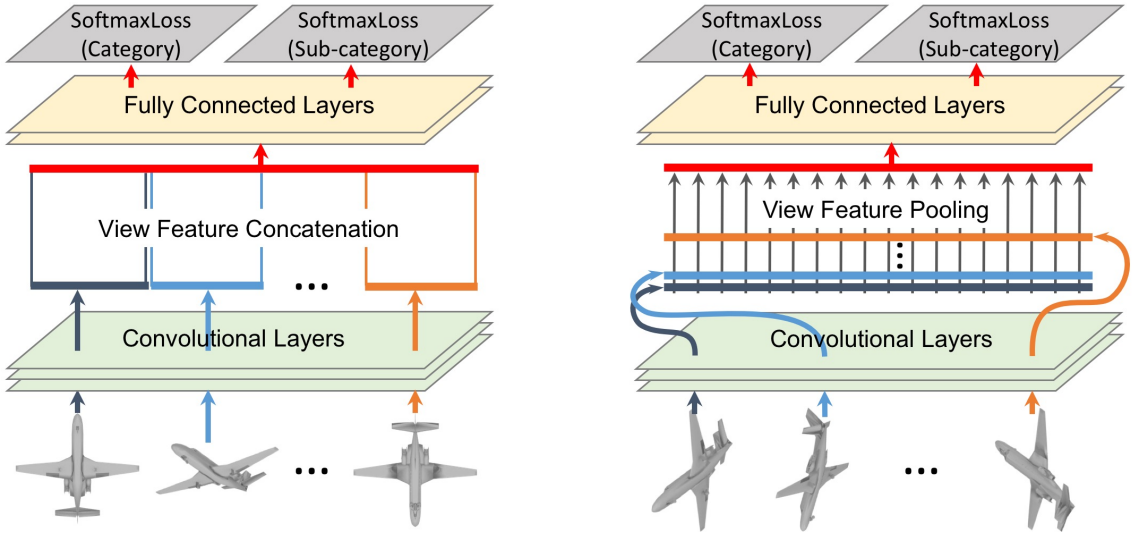

Figure 1: *CNN architectures for extracting 3D model signatures by aggregating CNN features of multi-view images rendered from 3D models. Multi-view CNN features are aggregated by concatenating per-view features when consistent 3D model orientation is available (left), or by max-pooling when model orientations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

1. Method Description

A 3D model can be rendered into a 2D image from multiple viewpoints, thus a possible signature for it can be obtained by an assembly of multi-view image features [CTSO03]. To generate a quality shape signature, image features must be informative and appropriately discriminative. In recent years, image features extracted from a CNN were shown to be highly successful in image-based recognition tasks, as they are both informative and discriminative. Considering the aforementioned multi-view render-based shape signature paradigm, such advances in image feature extraction can be levereged to boost the performance of shape signatures, for tasks such as shape retrieval.

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

In this approach, the aggregation of CNN features from different views is key. In [SMKLM15], the view features are aggregated via max-pooling, such that all views are equally contributive. This approach is therefore oblivious to the manner in which rendering views are chosen. This is an advantage when models within a collection are arbitrarily oriented, but a disadvantage when a consistent alignment is provided. An aggregation method that is consistency-aware is likely to outperform its oblivious counterpart in this instance. Thus, when a consistent shape alignment is given, instead of aggregation via max-pooling, we aggregate multi-view CNN features by concatenating view-specific features in-order. Here, the fully-connected layers can leverage the correspondence that exists between renders from consistent viewpoints across different shapes, and make decisions based on a more holistic understanding of the various viewpoints. See Figure 1 for a visual representation of the two CNN architectures employed in our approach.

More specifically, we render each model from 12 evenly spaced viewpoints, and fine-tune the VGG-19 network [SZ14] for the convolutional layers that are used for view feature extraction. We use the concatenation of the category and sub-category classification probability vector as the 3D model signature for retrieval. We open source our code at `http://yangyanli.github.io/SHREC2016/`.

References

[CTSO03] Chen D.-Y., Tian X.-P., Shen Y.-T., Ouhyoung M.: On visual similarity based 3d model retrieval. In *CGF* (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 *Proc. of ICCV* (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