3D CONVOLUTIONAL NEURAL NETWORKS BY MODAL FUSION

Yusuke Yoshiyasu, Eiichi Yoshida

AIST

Soeren Pirk, Leonidas Guibas

Stanford University

ABSTRACT

We propose multi-view and volumetric convolutional neural networks (ConvNets) for 3D shape recognition, which combines surface normal and height fields to capture local geometry and physical size of an object. This strategy helps distinguishing between objects with similar geometries but different sizes. This is especially useful for enhancing volumetric ConvNets and classifying 3D scans with insufficient surface details. Experimental results on CAD and real-world scan datasets showed that our technique outperforms previous approaches.

1. INTRODUCTION

Semantic object recognition from 3D geometry data is an important problem in computer vision, graphics and robotics, with the applications ranging from on-line shape search to autonomous robots. Due to the rapid growth of 3D big data, an accurate 3D shape search engine is highly expected. A good 3D recognition method is also the key to detecting objects from surrounding environment for autonomous robots, since 3D sensors are currently the most common choice for robotic vision.

Convolutional neural networks (ConvNets) have revolutionized the computer vision techniques in almost all the tasks: e.g., classification [10], detection [4, 5] and semantic segmentation [13, 29]. Further, by combining ConvNets with reinforcement learning, a computer can defeat professional human players in complex games such as Go [21] and a robot can learn to grasp an object by its own. ConvNets have also been becoming a powerful tool in 3D shape recognition—top-1 error evaluated on a 40-class classification task has been improved about 15% from hand-crafted features (75% → 90%).

What are the challenges and characteristics that are specific to 3D ConvNets, as compared to 2D ConvNets?

- **Input representation** Unlike 2D images, there is no decisive way of parameterizing an input for 3D ConvNets (image, volume, point, etc.), which needs explorations.
- **Fusion of multiple views and orientations** In contrast to 2D ConvNets, multiple features from different viewpoints or orientations are aggregated to construct a single descriptor for an object.
- **Encoding and fusion of modalities** Since we have 3D models in our hand, we can encode its geometry into various kinds of encodings i.e., geometric (e.g. surface normals), photometric (e.g. gray-level intensity) and physical (e.g. object size) encodings, etc. How to encode and to effectively fuse them needs explorations.

There are two main approaches to 3D ConvNets: the view-based approach that inputs multi-view images [23] and the volumetric approach that inputs 3D binary occupancy volumes [28]. The state-of-the-art results (classification accuracy) are around 90% and 85% for the view-based and volumetric approaches, respectively. Thus, there is still some room for improvements, especially for volumetric ConvNets. Recent works have tried to improve volumetric ConvNets using octree [26] or sub-volume supervision [19] to efficiently increase spatial resolution, but their performances have still not reached the level of view-based approach. [1]

In this paper, we explore how to encode input 3D models and to fuse them. In particular, we present multi-view and volumetric ConvNets that fuse geometric and physical information, i.e., surface normals and object size. To incorporate object size, we encode 3D models into height fields where a height value is assigned to each pixel and voxel. This strategy helps distinguishing between objects with similar geometries but different sizes, which provides an indirect yet effective way to improve the volumetric ConvNets. This is also effective for recognizing 3D scans where reliable features are difficult to be learned because of insufficient spatial resolution and noise. Experimental results on CAD model and scan datasets showed that our technique outperforms previous approaches.

2. RELATED WORK

3D ConvNets In 3D shape recognition field [20], ConvNets are becoming a powerful tool and has been improving the

[1] There are two common evaluation metric used in recent work: average *class* accuracy and average *instance* accuracy [19]. The readers should be careful about the metric used in the literature when comparing techniques. In this paper we use average *class* accuracy.

classification and retrieval performance. There are mainly two approaches to 3D ConvNets: volumetric and view-based approaches. Wu et al. proposed the first 3D ConvNets [28] using volumetric representation. Concurrently, Maturana et al. [15] also proposed an efficient volumetric approach called VoxNet for robot vision. Su et al. [23] proposed a multi-view technique which aggregates ConvNet features extracted from the images rendered from multiple views. Li et al. [11] introduced field probing neural networks (FPNN), a light weight 3D ConvNets for improving speed and reducing memory consumptions. Qi et al. [19] tried to close the performance gap between view-based and volumetric approach and found that augmenting azimuth rotations is crucial for improving accuracy of the volumetric technique. In recent works, diverse approaches of 3D ConvNets have been explored, such as 3D generative adversarial networks (3D GAN) [27] for unsupervised 3D shape recognition and 3D ConvNets for point clouds (PointNet) [18].

Modality fusion The fusion of multiple modalities has been studied in RGBD object recognition [24], detection [6], scene recognition [25] and action recognition. We further explore a fusion approach for 3D ConvNets.

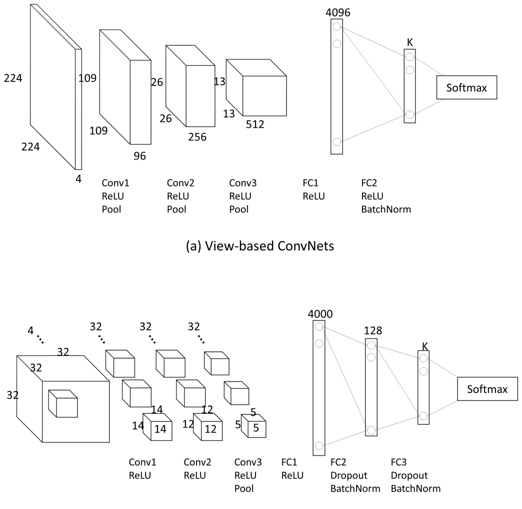
Object size in recognition While scale-invariance is a core property of shape descriptors for 2D images [14], a size of an object can play an important role, providing cues for recognizing objects. Hoeim et al. [7] used height information to detect objects in scenes. Frits et al. [2] showed that the absolute size of the object can augment image descriptors in recognition tasks. They checked the size of a bounding box to prune the erroneous predictions. In this paper we take an end-to-end training approach based on deep ConvNets to learn a fusion of local geometric features extracted from surface normal fields and object size information contained in height fields.

3. METHOD

We propose 3D ConvNets that combines different modalities. Here, we first explain our basic network architectures of view-based and volumetric ConvNets. We then describe the pooling methods designed for fusing modalities and viewpoints.

3.1. Basic network architecture

The basic network architecture excluding view and modality fusion layers are depicted in Fig. 1. We followed VoxNet [15] and used similar filter sizes and numbers to construct out volumetric network. The view-based network consists of 3 convolution layers and 2-3 fully connected (fc) layers, which is slightly shallower than AlexNet and VGG-M net for training and testing efficiency. This is also useful for achieving consistency between the two frameworks such that the comparison can be done easily. We insert dropout and batch normalization layers between fully connected layers, which are critical in avoiding over-fitting. We trained our networks from scratch without pre-training on other datasets.



(a) View-based ConvNets

(b) Volumetric ConvNets

Fig. 1. Basic network architectures (a) view-based. (b) volumetric.

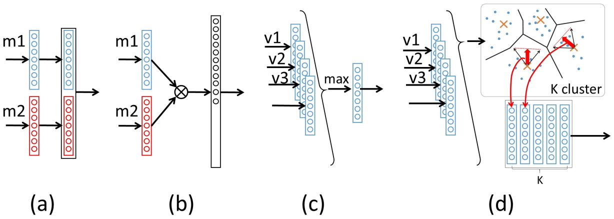


Fig. 2. Pooling methods. (a) concatenation, (b) bilinear, (c) max, (d) VLAD.

3.2. Feature map fusion techniques

Modality fusion When fusing different modalities, the goal is to learn the effective combinations of channels. For this, two types of pooling layers are considered: a concatenation and bilinear pooling [12] (Fig. 2 (a) and (b)). The simplest way to do so is to concatenate the modalities at the input layer. A more sophisticated way to do so is to use bilinear pooling (BP) which takes outer products of feature maps and captures multiplicative interactions between channels. The drawback of BP is that the resulting output features become high-dimensional. Thus, we use compact bilinear pooling (CBP) [3] that maps features to low dimensional space.

View and orientation fusion The simplest way to aggregate views is to take the average of softmax scores for all the views and then select the class with the maximum score (softmax summation). However, it is usually expected that better results can be obtained by a learning-based view aggregation method. In MVCNN, an average or max pooling layer is in-

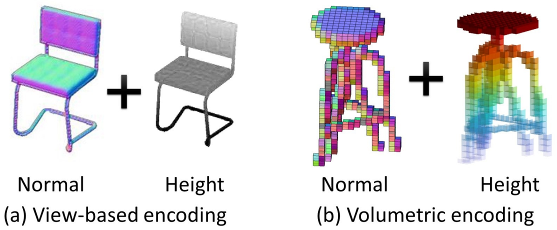


Fig. 3. Our geometric encoding that incorporates surface normal and height fields. (a) 2D view-based encoding. (b) 3D volumetric encoding.

serted into the network to aggregate views and trained end-to-end. Another order-less pooling technique is VLAD pooling technique [8], which is a variant of Bag-of visual word (BOVW). In recent work [1], the end-to-end CNN version of NetVLAD has been proposed. In this paper, we tested max pooling and NetVLAD for view pooling.

3.3. Input data encoding

In contrast to MVCNN that solely uses gray-level images or Voxnet that uses occupancy grids that contain 1 and -1, we incorporate physical size and local geometric information. Figure 3 shows our geometric encoding. We use surface normal and height fields for both 2D view-based and 3D volumetric methods. The proposed encoding is somewhat similar to HHA encoding that is used in the detection technique by Gupta et al. [6]. However, our focus here is to improve classification performances in 3D shape recognition. We have also tried several other encodings, such as depth and silhouette but we found that the combination of normal and height works effectively (Section 4, Table 2).

In order to generate the height encoding, we manually rescaled polygonal mesh models so that they conform with roughly their real-world size. For example, we rescaled chair models such that the height of seat become 450 mm, which is the standard chair height. We admit that the consideration of object size breaks scale-invariance, which is problematic in testing unknown 3D models in arbitrary size. However, recent CAD models that exist in the web are associated with size and thus we can easily rescale the model properly. Furthermore, we do not encounter this problem when testing 3D scans, as the metric is known. We also assume that the input 3D models are consistently aligned to an upright pose.

4. EXPERIMENTAL RESULTS

We implemented our algorithm on Matlab using Matconvnet toolbox. We used NVIDIA Geforce Titan X for training and testing. It takes approximately four hours to train our networks. Throughout experiments, we evaluate the classification accuracy based on *average class accuracy* [19, 23].

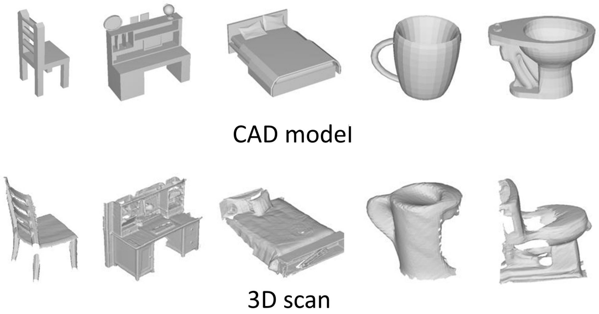


Fig. 4. 3D shape dataset.

4.1. Dataset

ModelNet We use ModelNet40 dataset [28] that contains CAD models from 40 categories for training and testing (Fig. 4 top).

Real-world scans We use the real-world scan dataset from Qi et al. [19], which comprises 243 scans from 12 categories. We use this dataset for testing only and training is done with Modelnet40. These scans are captured using a RGB-D sensor and densely reconstructed using the VoxelHashing framework (Fig. 4 bottom).

4.2. Experimental results

CAD models Fig. 5 (a) shows the classification results of our volumetric and view-based ConvNets using our encoding. We compared the proposed networks against Voxnets [15] that uses occupancy grids and MVCNN [23] that uses gray-level images. Our method using surface normal and height improves VoxNet by 6% (83.1% → **89.8**%) and MVCNN by 3% (90.1% → **93.3**%). Since the inputs of volumetric approaches tend to be low-resolution, the use of the height encoding is more effective in this case.

3D scans In Fig. 5 (b), we show the classification results of our technique on 3D scans. For both volumetric and view-based approaches, our algorithms improve the previous approaches by approx. 10%. In particular, our networks based on height fields helps distinguishing small and large objects with insufficient details, e.g., cups and toilets in Fig. 4. In fact, using height images that incorporate real-world scaling information, the average class accuracy of the cup model jumped from 0% to 61% (Table 1).

Other modalities We made an experiment by varying input modalities. Here we used gray-level intensity (I) which is obtained by rendering with the Phong lighting model, height map (H), depth (D), surface normal (N) and silhouette (S). For this experiment, view pooling is done with softmax summation and modality fusion is done with concatenation. In Table 2, we show the average class accuracies evaluated by varying modalities. With the use of height (H) only, it improves

	bathtub	bed	bench	chair	cup	desk	dresser	monitor	night stand	sofa	table	toilet	class ave.
MVCNN	100	100	95	82	0	88	75	93	14	69	39	65	68
Ours	86	100	100	71	89	69	25	91	81	73	94	100	**82**

Table 1. Classification results of 3D scan using view-based ConvNets.

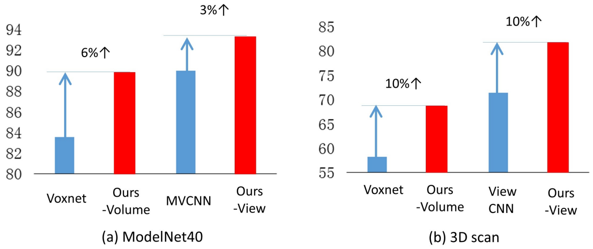


Fig. 5. Comparison with other state-of-the-art techniques.

the accuracy by approx. 3% from gray-level intensity (I). By combining surface normal and height (N + H), we obtained the accuracy of 91.25%, which is already beyond MVCNN, without using a learning-based view pooling technique.

Table 2. Comparison by varying encoding (View-based).

	I	N	H	D	S	N + H
Ave. Class Acc.	87.5	88.0	90.5	87.6	82.9	91.3

I–intensity; H–height; D–depth; N–surface normal; S–silhouette.

Table 3. Comparison of modality fusion (View-based).

	Concat (input)	Concat (conv3)	Concat (fc1)	CBP (fc1)
Ave. Class Acc.	91.25	91.12	92.37	91.13

Comparison between fusion techniques We compared the average classification accuracy of view-based ConvNets using different modality fusion techniques (concatenation and compact bilinear pooling (CBP)) for fusing surface normal and height. We also varied the place to insert a concatenation pooling layer (input, conv3, and fc1). Here view pooling was done based on a simple summation of softmax scores. As can be seen from Table 3, the concatenation fusion layer placed after the fc1 layer works best. Second, we compared the results obtained with different view pooling techniques (softmax summation, max and VLAD pooling). Here, modality fusion is done by concatenation at the input layer. We found that the max pooling works best for view pooling. For our setup (12 views and 2 modalities), a simple max pooling and concatenation works best for view and modality fusions. It would be interesting to test and compare the fusion techniques on a more complex setup with many modalities and views.

Comparison with size based pruning We compare our height encoding against the size-based pruning technique [2,22], which check the size of a bonding box. Specifically, if the size of the bounding box is out of 1-99 percentile of training dataset, it decreases the classification score of that object to zero. We combined this strategy with MVCNN as well as other unsupervised descriptors: d2 shape distributions [17], 3D Zernike moments [16] and spherical harmonics (SH) [9]. The average class accuracy of MVCNN + size pruning is 92% on CAD models, which is not as accurate as the proposed technique that learn to fuse surface normal and height. Furthermore, the average class accuracy only improved 4-5% from 71% to 75% for 3D scan classification.

Table 4. Comparison with size-based pruning technique.

	without size	with size
D2 descriptor	28.0	41.1
3D Zernike	34.8	45.5
Spherical Harmonics	64.0	70.5
MVCNN	90.0	92.3

5. CONCLUSION

We presented 3D ConvNets that fuse different modalities such as height and surface normals. This strategy was effective especially when the resolution of the input is low with insufficient details. The volumetric ConvNets was a good example where the resolution of input volumes cannot be easily increased due to computational demands. Further, this strategy was also effective for classifying 3D scans where reliable features are difficult to be learned because of insufficient spatial resolution of an acquisition system and noise. In future work, we will explore a way to automatically rescaling unknown CAD models.

6. REFERENCES

[1] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

[2] M. Fritz, K. Saenko, and T. Darrell. Size matters: Metric visual search constraints from monocular metadata. In J. D. Lafferty,

C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 622–630. Curran Associates, Inc., 2010. 2, 4

[3] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. *CoRR*, abs/1511.06062, 2015. 2

[4] R. Girshick. Fast R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015. 1

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[6] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from RGB-D images for object detection and segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 2, 3

[7] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *Int. J. Comput. Vision*, 80(1):3–15, Oct. 2008. 2

[8] H. Jgou, M. Douze, C. Schmid, and P. Prez. Aggregating local descriptors into a compact image representation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3304–3311, June 2010. 3

[9] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz. Rotation invariant spherical harmonic representation of 3D shape descriptors. In *Symposium on Geometry Processing*, June 2003. 4

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 1

[11] Y. Li, S. Pirk, H. Su, C. R. Qi, and L. J. Guibas. FPNN: field probing neural networks for 3d data. *CoRR*, abs/1605.06240, 2016. 2

[12] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *International Conference on Computer Vision (ICCV)*, 2015. 2

[13] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CVPR (to appear)*, Nov. 2015. 1

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004. 2

[15] D. Maturana and S. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, September 2015. 2, 3

[16] M. Novotni and R. Klein. 3d zernike descriptors for content based shape retrieval. In *The 8th ACM Symposium on Solid Modeling and Applications*, June 2003. 4

[17] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin. Shape distributions. *ACM Trans. Graph.*, 21(4):807–832, Oct. 2002. 4

[18] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CoRR*, abs/1612.00593, 2016. 2

[19] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2016. 1, 2, 3

[20] M. Savva, F. Yu, H. Su, M. Aono, B. Chen, D. Cohen-Or, W. Den, H. S. S. Bai, X. Bai, N. Fish, J. Han, E. Kalogerakis, E. G. Learned-Miller, Y. Li, M. Liao, S. Maji, A. Tatsuma, Y. Wang, N. Zhang, and Z. Zhou. Shrec16 track: Large-scale 3d shape retrieval from shapenet core55. In *In Eurographics Workshop on 3D Object Retrieval*, 2016. 1

[21] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503, 2016. 1

[22] S. Song and J. Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. 2016. 4

[23] H. Su, S. Maji, E. Kalogerakis, and E. G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proc. ICCV*, 2015. 1, 2, 3

[24] A. Wang, J. Cai, J. Lu, and T. J. Cham. Mmss: Multi-modal sharable and specific feature learning for rgb-d object recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1125–1133, Dec 2015. 2

[25] A. Wang, J. Cai, J. Lu, and T. J. Cham. Modality and component aware feature fusion for rgb-d scene classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5995–6004, June 2016. 2

[26] P.-S. Wang, Y. Liu, Y.-X. Guo, S. Chun-Yu, and X. Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions on Graphics (SIGGRAPH)*, 36(4), 2017. 1

[27] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Neural Information Processing Systems (NIPS)*, 2016. 2

[28] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1912–1920, 2015. 1, 2, 3

[29] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. In *International Conference on Computer Vision (ICCV)*, 2015. 1