PIEs: Pose Invariant Embeddings

Statistical Visual Computing Lab
UC San Diego

Chih-Hui Ho, Pedro Morgado, Amir Persekian, Nuno Vasconcelos
University of California, San Diego

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Introduction

- Pose invariant recognition is a difficult task, as an ideal embedding should map all the images of an object collected from multiple views into a single point.
- The introduction of multiview synthetic datasets, such as ModelNet[1], motivated a new wave of algorithms for multiview classification and retrieval.
- One of the most popular architectures is the multiview-CNN[2] (MVCNN), which complements a standard CNN embedding with a view pooling mechanism that produces a shape embedding.
- However, the multiview setting is not realistic for most real world applications, where there is no guaranteed that all the views will be available during test time.
- Previous works tend not to perform well for single view classification and retrieval, because the embedding of a single image (or view embedding) is not constrained to be similar to the shape embedding of its associated object.
- To overcome this issue, we propose pose invariant embedding (PIE) by encouraging
  - Different view embeddings from same object close to its shape embedding.
  - Different objects from same class close to its associated class embedding.
- Experiments show that PIE achieves good performance for both 1) classification and retrieval, and 2) single and multiview inference.
- The concept of PIE can generalize to CNN, triplet center[3] and proxy-NCA[4] based approaches, as illustrated in the taxonomy of embeddings in Figure 1.

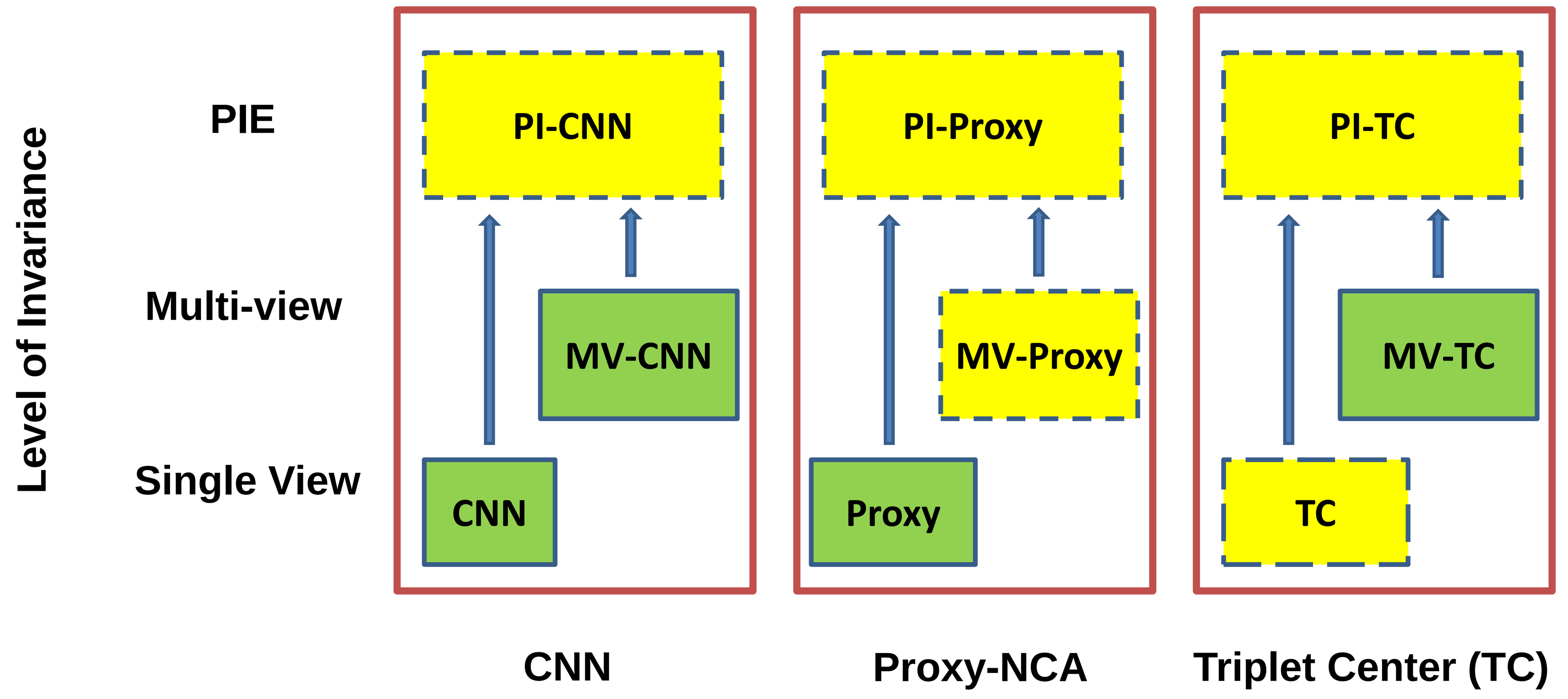
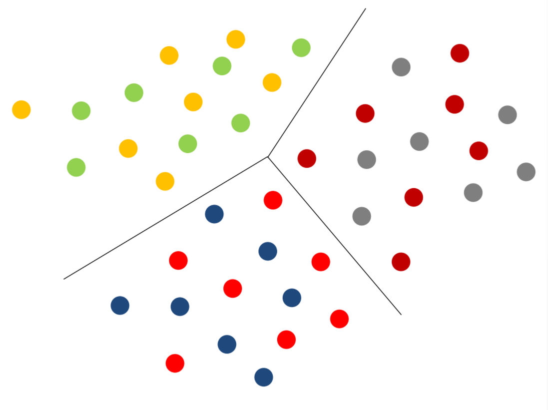


Figure 1. Taxonomy of embeddings learned by different methods according to different level of invariance. Green solid boxes represent methods in the literature and yellow dashed boxes represent methods proposed in this work.

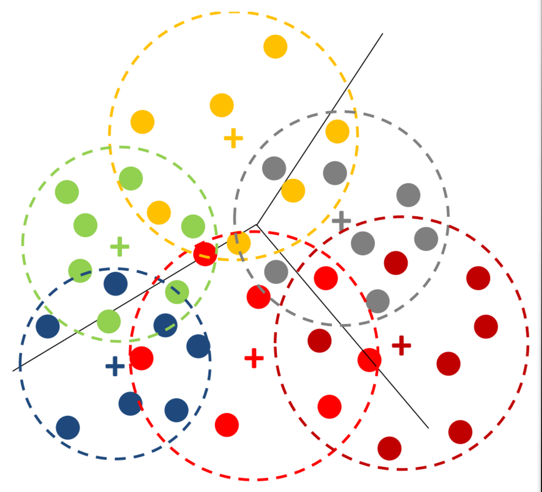
Proposed method

Embedding configuration	Description



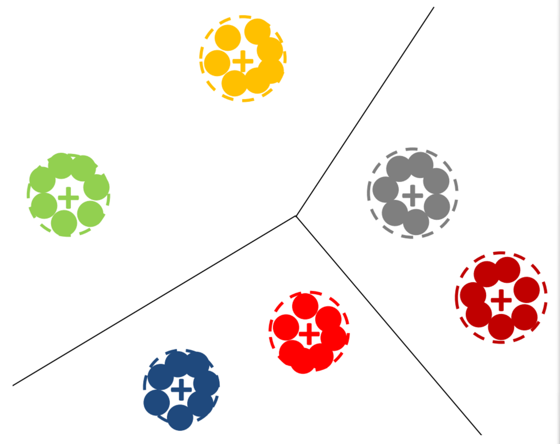
Single view

- Designed for single view task
- View embedding v of images from different objects but same class can interleave with each other
- Not a good embedding for tasks such as retrieving other views from same object
- Loss for proxy based network using single view embedding
- Loss = $\frac{\exp(-d(v,c_y))}{\sum_{i\neq y}\exp(-d(v,c_i))}$



Multiview

- Designed for multiview task
- Assume all views are provided during inference time
- No constraint between view embeddings v to its associated shape embedding s
- Performing worse on single view task
- Loss for proxy based network using multiview embedding
- Loss = $\frac{\exp(-d(s,c_y))}{\sum_{i\neq y}\exp(-d(s,c_i))}$



PIE (Ours)

- Applicable to both single view or multiview task
- Better embedding structure in embedding space
- Pose invariant distance is proposed for training
  - $d^{inv}(v,s,c_y) = \alpha * d(v,s) + \beta * d(s,c_y)$
- Loss for proxy based network using PIE
  - Loss = $\frac{\exp(-d^{inv}(v,s,c_y))}{\sum_{i\neq y}\exp(-d^{inv}(v,s,c_i))}$

● View embedding v

$d(.)$: Euclidean distance c: class embedding ✚ Shape embedding s

Dataset

A new multiview dataset, ObjectPI, is proposed for real world multiview task evaluation.
- Containing 500 real world objects
- Each object is imaged at 8 viewing angles
- Image contains complex background

Experiment

- 5 evaluation tasks on 3 datasets (ModelNet[1], Miro[5] and ObjectPI)
  - Classification (Cls.): single view cls., multiview cls.
  - Retrieval (Rtr.) : single view object rtr., single view class rtr., multiview class rtr.

Method	ModelNet (12 views) Classification (Accuracy %)			ModelNet (12 views) Retrieval (mAP %)			
	Single	Multi	Avg.	Object	Single	Multi	Avg.
RN[5]	80.2	89.0	84.6	22.6	20.2	63.9	35.6
MV-CNN[2]	71.0	87.9	79.4	29.6	41.7	71.5	47.6
PI-CNN	**85.4**	88.0	86.7	**50.8**	77.5	81.8	**70.0**
MV-TC[3]	77.3	88.9	83.1	36.6	63.5	84.0	61.4
PI-TC	81.2	88.9	85.1	41.4	71.5	84.2	65.7
MV-Proxy	79.7	**89.6**	84.7	35.0	66.1	**85.1**	62.1
PI-Proxy	85.1	88.7	**86.9**	40.6	**79.9**	**85.1**	68.6

Table 1. Comparison with state of the art multiview methods on ModelNet[1] dataset. Shadow denotes that the result of PIE is better than that of multiview based.

Table 2. Proxy based methods on ObjectPI. $\alpha = 1$, $\beta = 1$ is used in pose invariant distance for PI-Proxy.

Task		Proxy	MV-Proxy	PI-Proxy
Class. (Acc.)	Single	68.5	63.2	**68.7**
	Multi	78.8	78.3	**80.0**
	Avg	73.7	70.7	**74.4**
Retr. (mAP)	Object	47.7	49.3	**49.4**
	Single	59.7	57.9	**62.6**
	Multi	76.8	74.7	**78.2**
	Avg	61.4	60.6	**63.4**

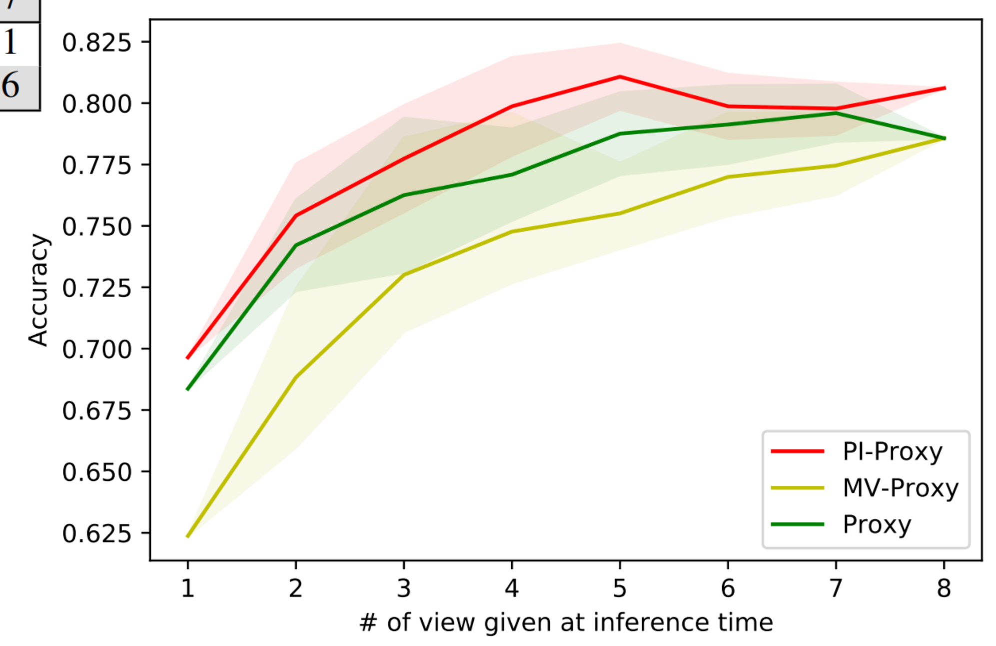


Figure 2. Classification accuracy (y axis) of ObjectPI as a function of number of views (x axis) given at inference time. PIE (red) is more robust to the number of views provided.

References

[1] Zhirong Wu, S. Song, A. Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1912–1920, June 2015.
[2] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In Proceedings of the IEEE international conference on computer vision, pages 945–953, 2015.
[3] Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. Triplet-center loss for multi-view 3d object retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 2018.
[4] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017
[5] Asako Kanezaki, Yasuyuki Matsushita, and Yoshifumi Nishida. Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018

Website & dataset available at svcl.ucsd.edu/projects/OOWL