STREET-TO-SHOP SHOE RETRIEVAL WITH MULTI-SCALE VIEWPOINT INVARIANT TRIPLET NETWORK

Huijing Zhan[1], *Boxin Shi*[2], *and Alex C. Kot*[1]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University
[2]Artificial Intelligence Research Center, National Institute of AIST
{hjzhan, eackot}@ntu.edu.sg, boxin.shi@aist.go.jp

ABSTRACT

In this paper we aim to find exactly the same shoes given a daily shoe photo (street scenario) that matches the online shop shoe photo (shop scenario). There are large visual differences between the street and shop scenario shoe images. To handle the discrepancy of different scenarios, we learn a feature embedding for shoes via a viewpoint-invariant triplet network, the feature activations of which reflect the inherent similarity between any two shoe images. Specifically, we propose a new loss function that minimizes the distances between images of the same shoes captured from different viewpoints. Moreover, we train the proposed triplet network at two different scales so that the representation of shoes incorporates different levels of invariance at different scales. To support training the multi-scale triplet networks, we collect a large dataset with shoe images from the daily life and online shopping websites. Experiments on the dataset show excellence over state-of-the-art approaches, which demonstrate the effectiveness of our proposed method.

Index Terms— Shoe Retrieval, Viewpoint Invariant, Triplet Network, Street-to-Shop, Feature Embedding

1. INTRODUCTION

Online shopping is getting more and more popular in these days and fashion related products purchasing contributes to a large portion of online sales. Driven by the huge profits, studies on fashion analysis have attracted increasing attention among computer vision researchers. There are a lot of works focusing on dealing with clothing items, which include clothing co-parsing (generating pixel-pise labels of the clothing items) [1, 2], clothing retrieval [3, 4, 5], clothing attribute classification [6, 7], etc. Though it is of comparable importance, the study on shoes [8] remains at the beginning stage.

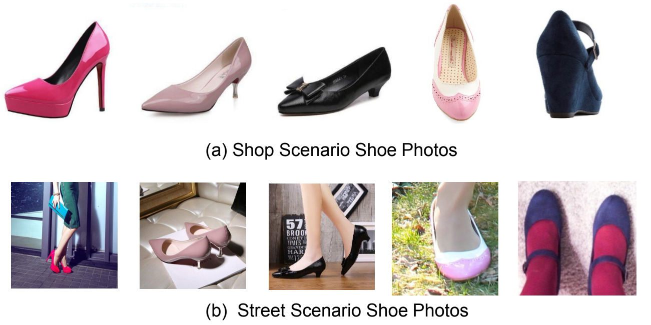
(a) Shop Scenario Shoe Photos
(b) Street Scenario Shoe Photos

Fig. 1. (a) Examples of shop scenario shoe photos. (b) Street scenario shoe images with cluttered backgrounds, scale and viewpoint variation, etc.

In this paper, we deal with a novel and practical problem: Given a real-world user photo depicting a shoe item, we find exactly the same item in online shops, defined as *street-to-shop shoe retrieval*. There are large visual discrepancies between shoes in the daily photos and online shop images as shown in Fig. 1. Firstly, online shop photos are usually captured in the white clean background, however, shoes in the daily photos have more diverse backgrounds. Secondly, online shoe photos are captured from several fixed poses while daily photos have more flexible poses. Lastly, there are large scale differences between shoes demonstrated in the online shoe photos and daily photos. These distinctive differences may lead to large Euclidean distances of even matched street and online shoe photo pairs, which makes the street-to-shop shoe retrieval task highly challenging. Therefore we learn a shoe-feature embedding so that the Euclidean distances of any two shoe images refers to their visual similarity.

Some recent works explored deep learning based approaches for feature embedding [9, 10, 11, 12, 13, 14] and achieved remarkable performances. Serra *et al.* [13] learned the style of clothes through a deep convolutional neural network (CNN) with triplet loss. A similar work to ours is [14] which proposed robust contrastive loss to learn the pairwise similarity for clothing images to facilitate cross-scenario

*This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Singapore, under its Interactive Digital Media (IDM) Strategic Research Programme. Boxin Shi is supported by a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

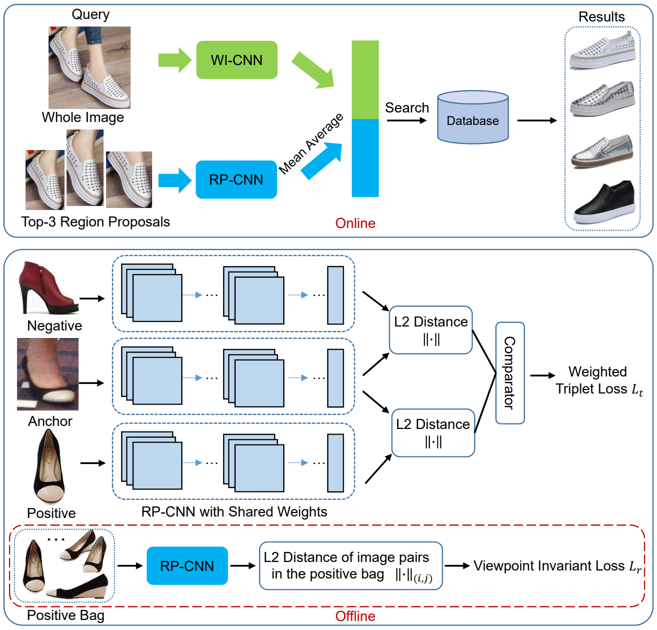

Fig. 2. The framework of our street-to-shop shoe retrieval system.

product search. However, to our best knowledge, no previous work specially considered the viewpoint invariance in the loss function, which is very important for the feature embedding for shoes.

The pipeline of our street-to-shop shoe retrieval framework is illustrated in Fig. 2. Firstly, we localize the shoes with high-quality region proposals to mitigate the negative effect of cluttered backgrounds. The deep feature embedding for shoes is learnt through convolutional neural network with the proposed new loss function. It integrates the viewpoint invariance in two aspects. On one hand, we assign higher weights to triplets from different scenarios compared with those from the same scenarios (weighted triplet loss L_t); On the other hand, images from different viewpoints of the same shoes (positive bag) are pulled together by minimizing their mutual feature distances (viewpoint invariant loss L_r). We then take the scale difference into account and independently train the multi-scale triplet networks with the whole images (WI-CNN) and high-quality region proposals (RP-CNN) for training. The query is represented as a concatenation of the features activated from RP-CNN and WI-CNN. Finally, the similarity is evaluated based on the L2 distance.

2. PROPOSED METHOD

In this section, we first describe triplet network followed by the proposed viewpoint invariant triplet network. Then we introduce our multi-scale triplet network training process with the novel region proposal selection method.

2.1. Triplet Network

Given a triplet of images $\{x_i^a, x_i^p, x_i^n\}$, x_i^a (anchor) and x_i^p (positive) have the same class labels, while x_i^n (negative) belongs to a different class. We wish to learn a deep embedding network $f(\cdot) \in \mathbb{R}^D$, which maps an input image x to a point in the D-dimensional Euclidean space. The goal of the function $f(\cdot)$ is to minimize Euclidean distances $d(f(x_i^a), f(x_i^p))$ of the matched pairs (x_i^a, x_i^p) while maximizing the distances $d(f(x_i^a), f(x_i^n))$ between the non-matched pairs (x_i^a, x_i^n). Inspired by [13], we normalize the feature distances to have the unit norm rather than the features. A softmax layer is built on top of the feature distances so that they are within the range of $[0, 1]$.

$$d_+ = \frac{\exp\left(d(f(x_i^a), f(x_i^p))\right)}{\exp\left(d(f(x_i^a), f(x_i^p))\right) + \exp\left(d(f(x_i^a), f(x_i^n))\right)}, \tag{1}$$

$$d_- = \frac{\exp\left(d(f(x_i^a), f(x_i^n))\right)}{\exp\left(d(f(x_i^a), f(x_i^p))\right) + \exp\left(d(f(x_i^a), f(x_i^n))\right)}. \tag{2}$$

Thus triplet loss L on a triplet is defined as:

$$L(x_i^p, x_i^a, x_i^n) = 0.5 \times [(1 - d_-)^2 + d_+^2] = d_+^2. \tag{3}$$

2.2. Viewpoint Invariant Triplet Loss Network

We incorporate the viewpoint invariance via a triplet weighting scheme in the training stage (weighted triplet loss) and an additional viewpoint invariant loss. Within a triplet of images denoted as $\{x_i^p, x_i^a, x_i^n\}$, they can be sampled from the same scenario or different scenarios. For example, assuming that (x_i^p) is sampled from the street scenario $(x_i^p)^s$ if (x_i^a) is from the online shop scenario $(x_i^a)^o$, then they are from different scenarios; otherwise, they are from the same scenario. Therefore all the triplets are classified into four situations: $\{(x_i^a)^s, (x_i^p)^o, (x_i^n)^o\}$, $\{(x_i^a)^s, (x_i^p)^s, (x_i^n)^s\}$, $\{(x_i^a)^o, (x_i^p)^s, (x_i^n)^s\}$ and $\{(x_i^a)^o, (x_i^p)^o, (x_i^n)^o\}$. We assume that the positive and negative images come from the same scenario. We set different weights to the loss incurred by the triplets from the same scenario and different scenarios by α and β, which are represented as L_s and L_d, respectively. Thus the weighted triplet loss $L_t(x_i^p, x_i^a, x_i^n)$ can be expressed as:

$$L_t(x_i^p, x_i^a, x_i^n) = \alpha L_s + \beta L_d. \tag{4}$$

Moreover, we also design a novel viewpoint invariant loss L_r. Even though the triplet network is viewpoint invariant in some aspects, we found in practice that it is more effective to complement it with an additional viewpoint invariant loss. Each shoe is depicted by $X^s = \{x_1^s, ..., x_{N_s}^s\}$, which contains N_s street shoe images, and $X^o = \{x_1^o, ..., x_{N_o}^o\}$ which contains N_o online shoe photos. Here we treat the online shoe

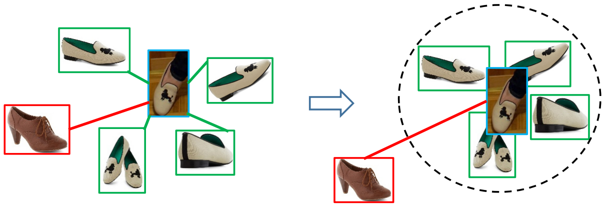

Fig. 3. The illustration for our viewpoint invariant triplet network. In the right figure, the positive examples of the same shoes are pulled closer with our viewpoint invariant triplet loss.

photos for the same shoes as a positive bag and make the features of the images in the positive bag similar to each other. This can be formulated to minimize the mutual distance of the image pair (x_i^o, x_j^o) in the online shoe photo set X^o, which can be written as:

$$L_r(X^o) = \frac{1}{2 \times n_d} \sum_{i,j}^{n_d} d^2(f(x_i^o), f(x_j^o)), \tag{5}$$

where n_d is the number of pairs in the set X^o. In our experiments, we randomly choose 5 image pairs. The concept of our viewpoint invariant triplet loss is illustrated in Fig. 3. The final loss J is composed of two parts:

$$J = L_t(x_i^p, x_i^a, x_i^n) + \lambda L_r(X^o), \tag{6}$$

where λ is the trade-off parameter. To further increase the network's invariance to viewpoint change, we also augment the data by rotating the online shop image in the range of $[-40°, +40°]$ with a step of 20°. We assume the rotation between the street image and online shoe image is a relative shift, thus the rotation is merely performed on the online shoe images. Moreover, to accelerate the speed of convergence, we carefully select the hard negative examples in the triplets. The feature representation for each shoe item X is represented as $f(X) = \frac{1}{N_o}\sum_{k=1}^{N_o} f(x_k^o)$. Every 5 epoches, the shoes are ranked in the descending order based on the L2 distances with the anchor of the triplets and we select the top 70% of them as the hard negative set of shoes.

2.3. Training the Multi-scale Triplet Networks

To make the representation of shoes scale invariant, we train the viewpoint invariant triplet network independently with two scales of images: the high-quality region proposals (fine scale) and the whole images (coarse scale). They are named as RP-CNN and WI-CNN, respectively. The training set for WI-CNN contains the street images with the whole image as the input. However, for RP-CNN, they are fed with the high-quality region proposals (usually contain an individual shoe).

The quality of the region proposals are evaluated in terms of three aspects: (1) the objectiveness score e returned by EdgeBox, (2) the probability score c from CNN detection model and (3) the overlapping score d with detections of DPM [15]. For the CNN detection model, it is trained as a binary classifier to differentiate whether a particular region proposal belongs to the foreground shoe or background.

We firstly employ EdgeBox [16] to produce an initial set of P ($P = 100$ in the experiments) region proposals and then develop a quality ranking scheme based on the rankSVM [17] for weights learning, the procedure of which is summarized in Algorithm 1.

Algorithm 1 Weights learning for the confidence scores

```
Input: An initial set of $P$ region proposals for image $I$;
Output: RankSVM weights $\mathbf{w}$;
for $j = 1$ to $P$ do
    Calculate the IoU[1]score $u_j$ with the annotated ground truth bounding box of image $I$;
    Forward the $j$-th region proposal into EdgeBox, CNN detection model and DPM model;
    return $\mathbf{h}_j = [e_j; c_j; d_j]$;
end for
Randomly sample $M$ pairs of region proposals denoted as $\mathbb{O} = \{(s_k, t_k)\}$, $k = \{1, 2, ..., M\}$ and calculate their pairwise relevance label $y_i(s_k, t_k)$;
$y_i(s_k, t_k) = \mathrm{sign}(u_{s_k} > u_{t_k})$;
Learn the weights $\mathbf{w}$ using the data pairs and their labels using rankSVM in Eq. (8);
return The weights of the confidence scores $\mathbf{w}$;
```

With the ordered pairs $\mathbb{O}$ and their pairwise labels y, each region proposal is represented by $\mathbf{h}$. Our goal is to learn a mapping function $S_f(\mathbf{h}) = \mathbf{w}^\top \mathbf{h}$ which predicts its quality score and estimates the relevance between data pairs (s_k, t_k) with the following constraint:

$$\forall u_{s_k} > u_{t_k} : S_f(\mathbf{h_{s_k}}) > S_f(\mathbf{h_{t_k}}). \tag{7}$$

The rankSVM model is built by minimizing the objective function:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{(s_k,t_k)\in\mathbb{O}} l(\mathbf{w}^T\mathbf{h_{s_k}} - \mathbf{w}^T\mathbf{h_{t_k}}), \tag{8}$$

where l is a loss function with the form $l(t) = \max(0, 1 - t)$ and C is the trade-off parameter. Given a query image q in the street scenario, the initially generated P region proposals are ranked in descending order according to the quality score $S_f(\cdot)$, and we choose top-3 scored candidates as the high-quality region proposals for RP-CNN network training.

[1] IoU is defined as the intersection of the region proposal window with the ground truth box divided by the union of them.

Table 1. Top-20 retrieval accuracy on our dataset.

Method	Top-20 Accuracy
VGG 16 Layers Pre-trained CNN [18]	35.28
Metric Network [9]	52.43
Triplet Embedding Network [13]	62.31
Robust Contrastive Loss [14]	61.20
WI-CNN Embedding Feature	65.45
RP-CNN Embedding Feature	67.90
Multi-Scale Viewpoint Invariant Triplet Network	**73.20**

3. EXPERIMENTS

3.1. Datasets

We collect about 14341 street scenario and 12652 shop scenario shoe images from two representative shoe shop websites, *Jingdong*[2] and *Amazon*[3]. Each shoe image has a unique ID according to its item number, which facilitates us to organize shoes with the same ID as matching pairs. A subset of shoe images in our dataset is annotated with bounding boxes. For evaluating the retrieval performance, 5021 daily shoe photos are used as the query and about 9500 online shop images are used in the reference gallery.

3.2. Implementation Details

We adopt the high performance VGG 16 Layers model [18] as our pre-trained network and implemented in Torch [19]. For optimization, SGD [20] is used and the mini-batch size is 16 triplets. The learning rate is initialized with 10^{-3}, momentum with 0.9 and weight decay of 10^{-4}. The image is resized to 256×256 and the center crop with size 224×224 is extracted. We set $\alpha = 1$, $\beta = 2$ and $\lambda = 0.001$ (to be of comparable magnitude) in our experiments. The trade-off parameter C is set as 0.001 in RankSVM. For the training of CNN detection model and DPM model [15], the positive examples are those cropped images which satisfy IoU > 0.8 and negative examples are thresholded with IoU < 0.2. The retrieval performance is evaluated based on the top-20 accuracy: If an exact match is found in the top 20 retrieved results, then it is regarded as a successful search; otherwise, it is a failure one.

3.3. Results on Our Shoe Dataset

We compare the proposed viewpoint invariant triplet network with several state-of-the-art techniques: (1) VGG 16 Layers Pre-trained CNN: Deep feature activated from the VGG 16 layers model with the whole image as input. (2) Metric Network [9]: A three-layer fully connected neural network with the deep feature activated from the AlexNet [21]. (3) Triplet Embedding Network [13]: The triplet network with the loss in [13]. (4) Robust Contrastive Loss [14]: An alternative of the contrastive loss with the siamese network. (5) WI-CNN Embedding Feature: The feature representation for the image is extracted from WI-CNN with the whole query image as the input. (6) RP-CNN Embedding Feature: RP-CNN activated deep features using the top-3 scored region proposals.

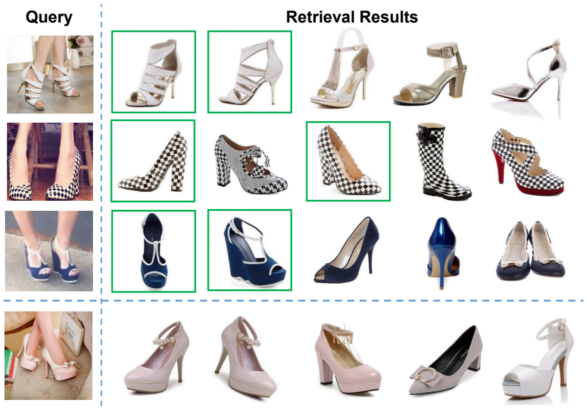

Fig. 4. Example retrieval results with the top-5 returned shoes. The exactly matched one is highlighted in green box.

Table 1 presents the top-20 retrieval accuracy on our shoe dataset, from which it can be seen that our proposed method outperforms the state-of-the-art methods. In particular, it outperforms VGG 16 Layers pre-trained CNN by twice. Compared with other deep architectures, it has a large improvement over [22] by about 11% and the siamese network [14] by about 12%. In comparison with the single-scale WI-CNN and RP-CNN embedding features, our multi-scale method increases about 8% and 5% respectively, which demonstrates the advantage of the multi-scale network.

Figure 4 illustrates the top-5 returned result on the dataset. The top three rows show the successful results and the bottom row is the failure case. It can be seen that our proposed method is capable of finding the exact same shoes in different viewpoints. Moreover, the false positives looks quite similar to the query. The failure example may be caused by the background clutter and the proposed method fails to capture the appearances for tiny area for the strap of the shoes.

4. CONCLUSION

In this paper, we address the street-to-shop shoe retrieval via a multi-scale view invariant triplet network, which embeds the feature of images for the same shoes similar to each other. We also assign different weights to the triplets from the same scenario or different scenarios. Experiments performed on our newly built dataset show its effectiveness. In the future, we plan to incorporate the semantic shoe attribute classification into the viewpoint invariant triplet network learning. For example, the attributes for heel shape are used as the supervised information to aid the learning of the triplet network.

[2] https://www.jd.com/

[3] https://www.amazon.com/

5. REFERENCES

[1] W. Yang, P. Luo, and L. Lin, "Clothing co-parsing by joint image segmentation and labeling," in *Proc. of Computer Vision and Pattern Recognition*, 2014.

[2] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Retrieving similar styles to parse clothing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 5, pp. 1028–1040, 2015.

[3] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan, "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set," in *Proc. of Computer Vision and Pattern Recognition*, 2012.

[4] J. Huang, R. S Feris, Q. Chen, and S. Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *Proc. of International Conference on Computer Vision*, 2015.

[5] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. of Computer Vision and Pattern Recognition*, 2016.

[6] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *Proc. of European Conference on Computer Vision*, 2012.

[7] A. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task cnn model for attribute prediction," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1949–1959, 2015.

[8] H. Zhan, B. Shi, and A. C. Kot, "Fashion analysis with a subordinate attribute classification network," in *IEEE International Conference on Multimedia and Expo*, 2017.

[9] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to buy it: Matching street clothing photos in online shops," in *Proc. of International Conference on Computer Vision*, 2015.

[10] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. of International Conference on Computer Vision*, 2006.

[11] S. Bell and K. Bala, "Learning visual similarity for product design with convolutional neural networks," *ACM Transactions on Graphics*, vol. 34, no. 4, pp. 98, 2015.

[12] Q. Yu, F. Liu, Y. Z. Song, T. Xiang, T. M. Hospedales, and C. C. Loy, "Sketch me that shoe," in *Proc. of Computer Vision and Pattern Recognition*, 2016.

[13] E. S. Serra and H. Ishikawa, "Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction," in *Proc. of Computer Vision and Pattern Recognition*, 2016.

[14] X. Wang, Z. Sun, W. Zhang, Y. Zhou, and Y. Jiang, "Matching user photos to online products with robust deep features," in *ACM Conference on Multimedia Retrieval*, 2016.

[15] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[16] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. of European Conference on Computer Vision*, 2014.

[17] Z. Cao, T. Qin, T. Liu, M. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *ACM International Conference on Machine Learning*, 2007.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[19] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, NIPS Workshop*, 2011.

[20] L. Bottou, "Stochastic gradient tricks," in *Neural Networks, Tricks of the Trade, Reloaded*, Lecture Notes in Computer Science, pp. 430–445. 2012.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.

[22] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. of Computer Vision and Pattern Recognition*, 2015.