Cross-Domain Shoe Retrieval using a Three-Level Deep Feature Representation

Huijing Zhan[†], Boxin Shi[‡], Alex C. Kot[†]

[†]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[‡]Artificial Intelligence Research Center, National Institute of AIST, Japan

***Abstract*—In this paper, we address the problem of matching the shoes from the daily life photos to exactly the same shoes from online shops. The problem is extremely challenging because of the significant visual differences between street domain images (shoe images captured in the daily life scenario) and online domain photos (images from online shops taken in the controlled environment). This paper presents a semantic Shoe Attribute-Guided Convolutional Neural Network (SAG-CNN) to extract the deep features. Moreover, we develop a three-level feature representation based on SAG-CNN. The deep features extracted from the image, region and part levels effectively match the images across different domains. We collect a novel shoe dataset, which consists of 8021 street domain and 5821 online domain images. The experimental results on our dataset show that the top-20 retrieval accuracy of our approach improves over that using the pre-trained CNN features by about 40%.**

I. Introduction

Online shopping explosively grows in the past few years, especially in the fashion domain with considerable sales. Particularly, the commerce on footwear rapidly increases with an annual growth rate of about 14% over the past two years [1]. In the daily life, we often come into such a situation, when seeing a nice pair of shoes on the shop window or worn on others' feet, we may want to find exactly the same pair from online shops with perhaps better prices. However, it is difficult for users to describe their desired shoes with several words to search through the text-based search engine. Thus a vision-based search system is required to search exactly the same shoes from online shops given daily life shoe photos. This problem is referred as the *cross-domain* shoe retrieval. We name the shoe photos captured in the daily life scenario as *street domain* while online shop pictures as *online domain*. Our problem is closely related with the topic of instance retrieval but differs in the sense that the query and reference images come from different domains.

An effective feature descriptor tailored for the object of interest is the key to a successful instance retrieval system. Existing systems adopt descriptors like GIST [2], Bag of Words (BoW) features from SIFT [3], Fisher Vectors from DSIFT [4], *etc*. However, these descriptors lose their effectiveness for cross-domain shoe retrieval, due to the unique challenges such as viewpoint variation, scale variation, cluttered background, and self-occlusion as shown in Fig. 1. A recent work in [5] proposed a metric network to evaluate the similarity of feature vectors from different domains. However, the feature vectors are generated from the FC1 layer of the pre-trained AlexNet [6] on the whole image, which are not sufficient to capture the appearance of shoes in different scales. To deal with the large visual differences of different domains, it is essential to develop a multi-level discriminative shoe feature representation for minimizing the distances of the same shoes from different domains, while maximizing the distances for different shoes.

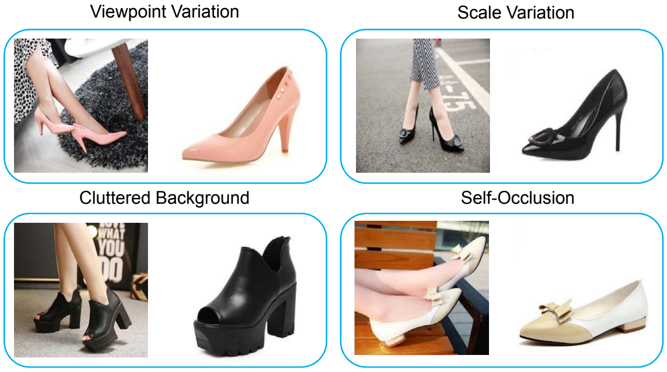

Fig. 1. The visual differences between street domain photos and their counterpart online domain images are illustrated. In each blue rectangle, the left image is from the street domain and the right image shows exactly the same individual shoe from the online domain.

In this paper, we deal with cross-domain challenges from scale variation and cluttered background. Specifically, we incorporate semantic shoe attribute-guided convolutional neural network (SAG-CNN) for discriminative feature learning and develop a three-level feature representation for shoes with three scales of input to SAG-CNN. The first level is the whole image, the second level is the top-1 selected region proposal using our newly proposed region proposal selection algorithm, and the third level corresponds to the shoe part image patches detected by Deformable Part Model (DPM) [7]. The three-level feature representation describes the appearances of shoes in terms of the global structure information and local patch details. We evaluate the performance of our system on a newly built cross-domain shoe dataset, which is able to achieve a good retrieval accuracy.

*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.

978-1-4673-6853-7/17/$31.00 ©2017 IEEE

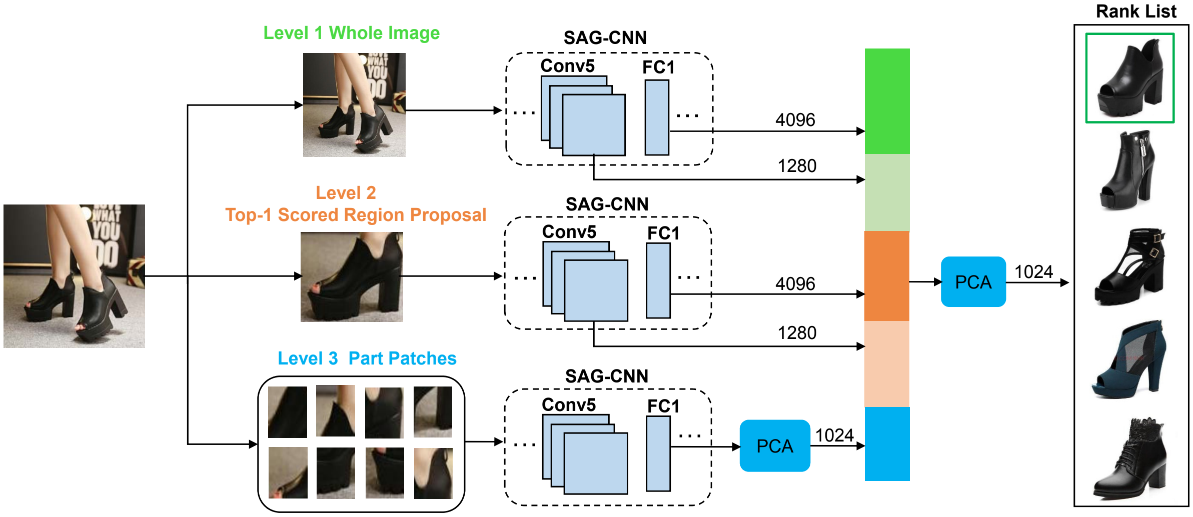

Fig. 2. The pipeline of our proposed shoe retrieval system.

II. Dataset Construction

We collect about 8021 daily shoe photos from the street domain and 5821 shoe pictures from the online domain. Each shoe image has a unique ID according to its shoe model number, which facilitates us to organize shoes with the same ID as matching pairs. Each online shoe photo contains an individual shoe facing in the left side 45 degree view and clean background. The images are crawled from *Jingdong*[1] with each shoe model accompanied with several daily life photos and its corresponding online domain shoe photos.

The attributes of shoes (*e.g.*, toe shape, color, pattern, style, *etc.*) are extracted from text descriptions next to the crawled images. We then manually re-organize the attribute using the following three procedures: 1) Remove attribute types that are not visible from the image such as sole material, *etc.*; 2) Re-annotate the crucial attributes that determine the shoe's appearance such as heel type, toe shape and color; 3) Merge the attribute values that are semantically close. For example, "beige" and "ivory teeth" can be considered as one attribute named "white". In total, we have 11 types of semantic attributes with 121 values. To our best knowledge, we have established the first shoe image dataset with semantic annotations for corresponding images from street and online domains.

III. The Proposed System

The pipeline of our proposed shoe retrieval framework is illustrated in Fig. 2. Given a query image in the street domain, we represent it using three-level features activated from the SAG-CNN network with three different scales of images as input: the whole image (Level 1); the top-1 scored region proposal (Level 2), which highly likely contains the shoes; part patches detected by DPM (Level 3). For reference images from the online domain, the representation almost follows the same pipeline; the only difference is that the Level 2 image is replaced by the whole image because of the white background. The distances of feature vectors between query image and reference images are evaluated in terms of cosine similarity.

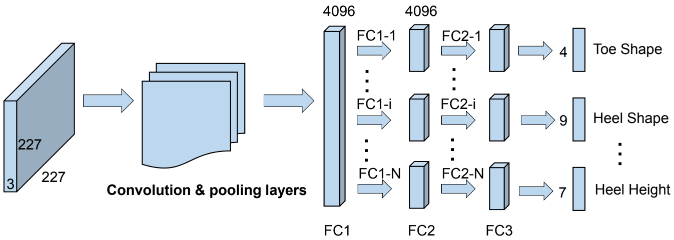

Fig. 3. The architecture of our SAG-CNN network. It consists of five convolution layers and one fully connected layers shared by all the attributes.

A. Network Architecture of SAG-CNN

We employ the convolutional neural network for discriminative shoe feature learning. The network structure is illustrated in Fig. 3. We make appropriate adaption based on the AlexNet [6] by adding a tree-structure set of fully connected layers after FC1 layer to obtain attribute-sensitive features. For each attribute, its last fully connected layer branches out several units, according to the number of the attribute values. For example, the attribute like "toe shape" can take 4 values then its specific fully connected layer branches out 4 units, indicating the probability of having the corresponding attribute value.

B. Three-Level Feature Representation For Shoes

To address the cross-domain visual discrepancy caused by the scale variation and background clutter, a discriminative three-level feature representation is designed for shoes. We begin by introducing how we generate the Level 2 feature using a newly proposed region selection approach and introduce briefly about the Level 1 and Level 3 features later, which are trivial to obtain.

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

1) Level 2 feature: We model the task of localizing the shoes with a bounding box as a high-quality region proposal selection procedure. State-of-the-art region proposal generation approaches like Selective Search [8] and EdgeBox [9] produce an initial set of region proposals, which contains many noisy candidates with low Intersection-Over-Union (IoU) scores. IoU is defined as the intersection of the region proposal window with the ground truth box divided by the union of them. Therefore, it is essential to develop a ranking strategy to re-rank the initial pool of region proposals.

We use three criteria to evaluate the quality of the region proposals: the objectiveness score returned by EdgeBox (e), the probability score from CNN detection model (c) and the confidence score by DPM (d). Then rankSVM [10] is employed to learn the weights measuring the importance of three scores. 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.

Algorithm 1 Weights learning process of the confidence scores

```
Input: A set of $N$ training images from the street domain;
Output: RankSVM weights $\mathbf{w}$;
for $i = 1$ to $N$ do
    Generate an initial pool of $P$ region proposals using EdgeBox;
    for $j = 1$ to $P$ do
        Calculate the IoU score $u_j$ with the annotated ground truth bounding box of image $I_i$;
        Forward the $i$-th region proposal into EdgeBox, CNN detection model and DPM model;
        $\mathbf{h}_j = [e_j; c_j; d_j]$;
    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 $y_i(s_k, t_k)$;
    if $u_{s_k} > u_{t_k}$ then
        $y_i(s_k, t_k) = 1$;
    else
        $y_i(s_k, t_k) = -1$;
    end if
end for
Learn the weights $\mathbf{w}$ using the data pairs and their labels using rankSVM [10];
return The weights of the confidence scores $\mathbf{w}$;
```

The weights learning process is demonstrated in Algorithm 1. 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 $f(\mathbf{h}) = \mathbf{w}^\top \mathbf{h}$ which predicts its corresponding quality score. It should estimate the relevance relationship between data pairs (s_k, t_k) with the following constraint:

$$f(\mathbf{h}_{s_k}) > f(\mathbf{h}_{t_k}), \text{if } u_{s_k} > u_{t_k}, \tag{1}$$

where u_{s_k} and u_{t_k} are IoU scores of the region proposal pair (s_k, t_k) with the ground truth boxes. The rankSVM model is built by minimizing the objective function:

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

where ℓ is a loss function with the form $\ell(t) = \max(0, 1-t)$ and C is the trade-off parameter.

Given a query image q in the street domain, it can be represented as $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_P]$ with the confidence scores computed from the P initially produced region proposals. Then the quality scores are denoted as $\mathbf{J} = \mathbf{w}^\top\mathbf{H}$, where $\mathbf{J}$ is P-dimensional vector, with each element indicating the quality score for the corresponding region proposal. We choose top-1 scored region proposal as the Level-2 image, the activation of which from the Conv5 and FC1 layers of the SAG-CNN is represented as $F_{L2} \in \mathbb{R}^{5376}$. Here we apply the two level pyramid mean-pooling [11] $(2\times2, 1\times1)$ on the Conv5 feature map.

2) Level 1 and Level 3 features: For the Level 1 feature, we directly feed the global image into the SAG-CNN and the activations from the Conv5 and FC1 layers are used for representation, denoted as $F_{L1} \in \mathbb{R}^{5376}$. The fine-grained parts of the shoes detected by DPM are fed forward into FC1 layer of the SAG-CNN and features are concatenated as the final representation for the Level 3 feature. The PCA is used to reduce the dimension of the Level 3 feature F_{L3} to 1024-D. The resulting concatenated three-level feature is a $5376 + 5376 + 1024 = 11776$-D vector, which is further reduced to 1024-D.

IV. Experimental Results And Discussions

A. Experimental Settings

We train different SAG-CNN models for the street and online domains separately because the appearances of the shoe attributes show more variances in the street domain than in the online domain, due to the viewpoint change, *etc*. We produce about 15000 top-5 scored region proposals of 3000 shoe images for street domain SAG-CNN model learning and 1800 clean shoe photos in the online domain SAG-CNN learning, which are increased to 18000 images after data augmentation.

Next we introduce the training of models involved in the Level 2 feature generation process. The EdgeBox algorithm with default parameters is used to generate $P = 100$ initial region proposals. About 2600 shoe images are labeled with bounding boxes. For the CNN detection model, we generate about 42150 cropped images with IoU > 0.8 as the positive samples and 67440 cropped negative images with IoU < 0.2. For the DPM detection model, about 500 ground truth annotated shoe images as the positive and 2000 negative cropped images with IoU < 0.2 are used to train a 5-component DPM model. Note that the learnt DPM model is also used to detect the shoe image patches of the third level. Finally, we randomly choose 100 shoe images to generate the ordered pairs for weights learning in RankSVM.

For evaluating the retrieval performance, 5021 daily shoe photos are used as the query with each one having the corresponding shoes in the reference set and the rest 4021 online domain images are used as the reference gallery.

Table I
TOP-20 RETRIEVAL ACCURACY OF OUR SYSTEM AND BASELINE RESULT

Method	**Top-20 Accuracy**
Gist feature [2]	11.13
DSIFT + Fisher Vector [4]	20.04
Deep feature (FC1) of Pre-trained CNN	28.28
Deep feature (FC1) of SAG-CNN	44.67
Metric Network [5]	52.43
Three-level feature (FC1 + Conv5) with SAG-CNN	**66.92**

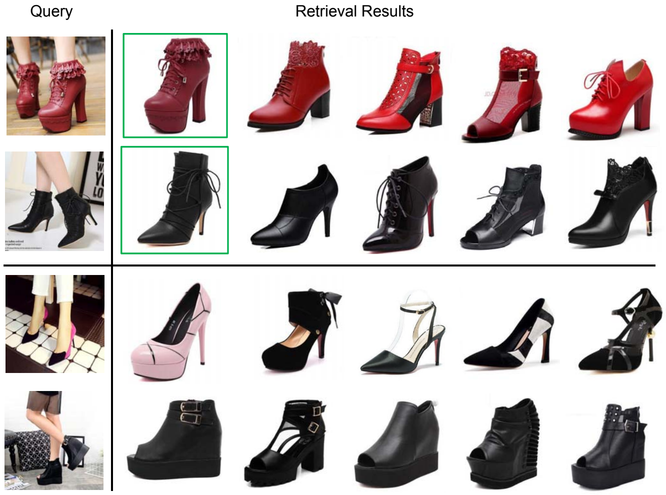

Fig. 4. Example retrieval results with top-5 returned shoes are shown.

B. Performance Evaluation and Comparison

The retrieval performance of our proposed system is evaluated in terms of the top-20 retrieval accuracy. If the top-20 returned results contain exactly the same individual shoe to the query item, then it is considered as a successful result; otherwise, it is a failure case. To quantitatively validate the effectiveness of our proposed three-level SAG-CNN activated feature, we adopt the following baselines for comparison: 1) *GIST feature* [2] with 512-dimension and Dense SIFT feature followed by fisher vector encoding (*DSIFT + Fisher Vector*) [4] with the codebook size $D = 64$; 2) *Deep feature (FC1) of Pre-trained CNN* : deep feature activated from the FC1 layer of the pre-trained AlexNet [6] with the whole image as the input; 3) *Deep feature (FC1) of SAG-CNN*: deep feature extracted from the FC1 layer of the SAG-CNN network with the whole image as the input; 4) *Metric Network* [5], which is used to evaluate the similarity of feature vectors from different domains. The feature vectors are generated from the FC1 layer of the pre-trained AlexNet on the whole image.

According to the experiment result shown in Table. I, the deep feature activated from pre-trained AlexNet outperforms the state-of-the-art system utilizing the DSIFT + Fisher Vector by about 8%. The top-20 accuracy using the Deep FC1 feature extracted from SAG-CNN network improves about 16% compared with that using the pre-trained AlexNet model, which demonstrates the effectiveness of the SAG-CNN network guided by semantic shoe attribute learning. We also compare the performance of our proposed system with the recent work on cross-domain product retrieval [5]. The retrieval accuracy of our proposed system improves about 14% over their method.

Example retrieval results are shown in Fig. 4. The top 2 rows indicate the successful retrieval results while the last two rows show the failure examples. The successful retrieval examples show that our system is capable of handling some challenging pose like the 1^{st} case, also the background clutter (*e.g.*, the poster beside the high-heel shoes in the 2^{nd} row). For the failure examples, we notice they are caused by the complex patterns of the background (*e.g.*, the patterns of the floor look similar as the stripe of shoes in the 3^{rd} row). Also it is still difficult for our proposed feature representation to capture the subtle decoration details (*e.g.*, the number of buckles in the query image in the 4^{th} row).

V. CONCLUSIONS

In this paper, we present a shoe retrieval system which aims at finding the exact same shoe image in online shops according to query image from the daily life. In addition, a novel shoe dataset is introduced, which consists of street domain and online domain matching pairs with fine-grained semantic attributes. We develop a discriminative three-level deep feature representation extracted from the SAG-CNN, which achieves substantial improvement over deep feature from the pre-trained AlexNet and outperforms recently proposed method on product retrieval.

REFERENCES

[1] Insights in GetFriday, "A study of the online shoe retailer industry," 2015.

[2] Aude Oliva and Antonio Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[3] Josef Sivic and Andrew Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2003, pp. 1470–1477.

[4] Herve Jegou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Perez, and Cordelia Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1704–1716, 2012.

[5] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg, "Where to buy it: Matching street clothing photos in online shops," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3343–3351.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[7] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva 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.

[8] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[9] C Lawrence Zitnick and Piotr Dollár, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision*. Springer, 2014, pp. 391–405.

[10] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li, "Learning to rank: from pairwise approach to listwise approach," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 129–136.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision*. Springer, 2014, pp. 346–361.