Tagging the Shoe Images by Semantic Attributes

Huijing Zhan, Sheng Li, Alex C. Kot

Rapid-Rich Object Search(ROSE) Lab
School of Electrical and Electronic Engineering
Nanyang Technological University
Singapore 639798

***Abstract*—With the rapid proliferation of Internet, it becomes a great challenge to annotate explosive number of objects manually. Especially for the fashion domain where a massive collection of new products come up everyday. Therefore, to save human labor, it is essential to develop an automatic tagging system for those fashion products in a variety of appearances. In this paper, we focus on addressing the issue of automatic shoe tagging where a novel system is proposed to predict the semantic attributes of the shoe images. Given a shoe image in an unknown viewpoint, our proposed system first classify it into one of the 6 predefined representative viewpoints, which are commonly displayed in online merchants. To localize the shoe parts on the identified viewpoint, a view-specific part localization model is proposed based on the prior knowledge of the shoe structures under different viewpoints. Finally, we extract several complementary low-level features from the localized shoe parts, which is fed into a SVM classifier for attribute prediction. The effectiveness of the proposed system is demonstrated on a newly-built pump shoe image dataset.**

***Index Terms*—shoe tagging; view classification; view-specific part localization; attribute prediction**

I. Introduction

With the explosive growth of Internet which brings ever increasing visual data, it is urgent to develop techniques to tag these data for a good search experience, especially for the online merchants selling fashion items, where a great many new products are uploaded every day. In recent years, researchers have been using the "attribute" for efficient tagging, which is a mid-level concept targeting to bridge the interpretation gap between machine and human. The attribute-based tagging techniques have been applied on the search of fashion items, such as clothes [1][2][3][4], bags [5] or shoes [6][7][8].

Among the works for the fashion item search, the studies on image based shoe retrieval is still at its infancy. Kovashka *et al.* [6] develops a shoe retrieval system using relative attribute based on comparisons. It is capable of whittling away irrelevant shoes through interactive users' feedbacks. However it is difficult to get a consensus on relevance feedbacks among annotators, resulting to a human preference-biased comparison. To address this issue, a new concept named "adaptive attributes" is proposed in [7], which adapts the general attribute function to a user-specific one. However, the users' taste is inconsistent and difficult to be modeled systematically. The work in [8] proposes a shoe retrieval framework via part-related attributes. It jointly optimizes shoe part detection and attribute prediction based on Deformable Part Model(DPM) [9]. Localizing the shoe part is time consuming due to the use of DPM [10]. What is more, it employs a set of small bounding boxes to represent each part, which may not contain sufficient discriminative information. On the other hand, SIFT features might be not that appropriate for the shoe attribute prediction task. Empirically, we find that the detected SIFT keypoints is not only sparse but also could not reflect the shape details of the shoe designs. Thus several complementary features are extracted from a shoe image to better capture the shoe's appearance, from both the global and local points of view.

In this paper, we focus on semantic tagging of the shoe images for online merchants, so as to facilitate the task of shoe image retrieval and shoe indexing. We propose a novel shoe image tagging system to predict the semantic attributes of shoe images. Altogether, we define seven part-aware shoe attributes which are binary. Firstly, a view identification stage is employed to classify the viewpoint of the input shoe image. Specifically, we identify it into one of the 6 representative viewpoints which are mainly displayed in online merchants. Then, for each shoe attribute (termed as attribute for short), our system predicts its value based on the identified viewpoint. During the attribute prediction, a novel view-specific part localization method is proposed to localize the shoe parts (which accommodates the attribute) on the identified view. Complementary features are then extracted from the localized parts and concatenated for the final prediction. We evaluate the performance of our system on a newly built pump shoe dataset, which is able to achieve a good attribute prediction accuracy.

II. Pump Shoe Dataset

The images of our pump shoe dataset are collected from Amazon.com, where the shoes are usually displayed close to 6 representative viewpoints as shown in Fig. 1. These shoe images are usually captured by the professions with clean background. In total, our pump shoe dataset consists of 7500 shoe images with 1250 shoes in each of the 6 viewpoints.

For each shoe, the groundtruth annotation contains seven binary values corresponding to seven part-aware attributes which are defined based on four different shoe parts, as shown in Table I. While the groundtruth annotation of each attribute is collected manually.

978-1-4799-8058-1/15/$31.00

Fig. 1. The six representative viewpoints of a shoe

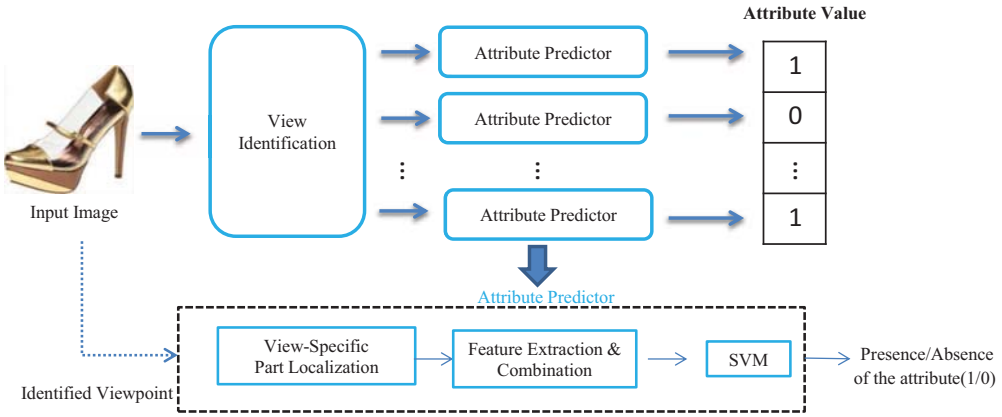

Fig. 2. The proposed system

TABLE I
PART-SPECIFIC ATTRIBUTES AND THE CORRESPONDING RELEVANT VIEWS

Shoe Parts	Related Attributes
head	closed toe, pointy
body	side-covered, bounds
back	back-covered
heel	high thin heel, wedge heel

III. The Proposed System

Fig. 2 illustrates the overall framework of our proposed system for the shoe image tagging. Given an input shoe image, we first classify which viewpoint it belongs to or is close to. Then, the shoe image is fed into a set of attribute predictors to estimate the attribute values. For each attribute predictor, it localizes the shoe part that accommodates the corresponding shoe attribute, which is done by using our proposed view-specific part localization model. For the localized parts, proper complementary features are extracted and combined together for predict the part-aware attribute value.

A. Viewpoint Identification

Most of the product shoe images sold by online merchants are displayed in or near to the above defined 6 viewpoints as shown in Fig. 1. These 6 poses are captured in the left-profile (1^{st} viewpoint), frontal (2^{nd} viewpoint), back (3^{rd} viewpoint), left (4^{th} viewpoint) and right facing viewpoint (5^{th} viewpoint) as well as the top view (6^{th} viewpoint). We pre-train a SVM viewpoint classifier by using the GIST features. Given a input shoe image, we extract its GIST features and feed it into the viewpoint classifier to identify the viewpoint.

B. View-specific Part Localization

In this section, we propose a novel shoe part localization method which is view-specific. Our proposal is based on the observation that the appearances for the shoe parts are pretty much structured for a certain view.

First of all, we obtain the shoe region which is a rectangular by removing the background. This is a trivial preprocessing since most of the shoe images from online merchants are clean. In the following discussions, a shoe image refers to the one after such background removal.

We firstly learn a view-specific part localization model for each of the 6 representative viewpoints. Let $I = \{I_i^v\}_{i=1}^N$ denote the N shoes under the v viewpoint for training the part localization. For each shoe image I_i^v, at most four parts are annotated as bounding boxes as illustrated in Fig.3. We define the annotated bounding box as $B_{ip}^v = [B_{ip}^v(x_1), B_{ip}^v(x_2), B_{ip}^v(y_1), B_{ip}^v(y_2)]$, where $p = 1, 2, 3, 4$ refers to the respective parts of a shoe, $B_{ip}^v(x_1)$, $B_{ip}^v(x_2)$ and $B_{ip}^v(y_1)$, $B_{ip}^v(y_2)$ indicate two horizontal and vertical coordinates of the bounding box, respectively. Note that some parts might be not available for a certain view. Such parts will not be considered for part localization in this view. The bounding box B_{ip}^v are further normalized by

$$\hat{B_{ip}^v} = B_{ip}^v ./ T_i^v \quad (1)$$

where $T_i^v = [w_i^v, w_i^v, h_i^v, h_i^v]$ is a vector records the width w_i^v and height h_i^v of I_i^v, and "./" denotes the element wise dividing. We obtain the following part localization model for

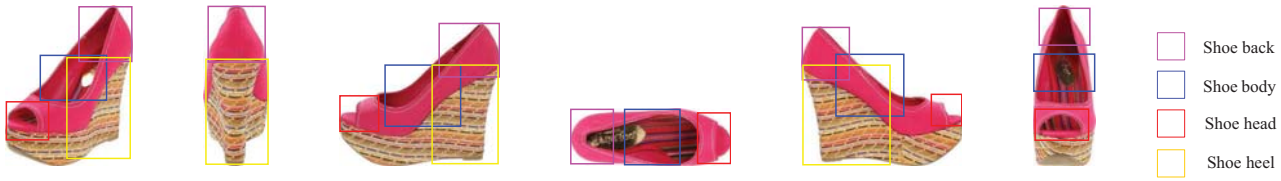

Fig. 3. Examples of manually annotated parts for shoe images of 6 representative views(best viewed in pdf file and color images)

each of the 6 representative viewpoints v by

$$\bar{B}_p^v = \frac{\sum_{i=1}^{N} \hat{B}_{ip}^v}{N} \quad (2)$$

Given a test image I'^v with the identified viewpoint v, its shoe part can be localized as

$$B_p'^v = \bar{B}_p^v . \times T'^v \quad (3)$$

where $T'^v = [w'^v, w'^v, h'^v, h'^v]$ is a vector records the width w'^v and height h'^v of I'^v, and ".×" denotes the element wise multiplication.

C. Feature Extraction and Combination

For the attribute prediction, we extract three types of complementary features on the localized parts on the given view, including Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), GIST. LBP features are commonly used to analyze the texture patterns of the image, while HOG features are better suited for the task of evaluating the edge information. Compared with LBP and HOG which are local based features, GIST features are capable of characterizing and describing the shape of the objects. Therefore, these three modes of features work in a complementary way providing not only rough but also fine-detailed visual concepts of shoes. Finally, the combined feature will be used for predicting the attribute value using a SVM classifier.

IV. Experimental Results And Discussions

A. Experimental Settings

We evaluate the effectiveness of our proposed system on our newly built shoe dataset. Preliminarily, each image is normalized to 500 pixels at maximum for side length. For the data preparation at the training and testing phase, we use the following settings. We partition the 7500 images in our dataset into the following two non-overlapping parts: 1)1500 images (250 images for each of the 6 viewpoints) for the training of viewpoint classifier and attribute predictor (including the view-specific part localization model and SVM classifier for attribute value prediction), and 2) the rest 4500 images are used for testing. Five fold cross-validation scheme is carried to determine the parameters of the SVM classifiers.

B. Performance Evaluation and Comparison

We fist evaluate the performance of the viewpoint identification, GIST features with the dimension equal to 3200 are extracted for both the training and testing of the viewpoint classifier, which is able to achieve 98.3% viewpoint identification accuracy.

The performance of our proposed system is evaluated based on the attribute prediction accuracy. We compare the performance of our system with the state-of-the-art [8]. For a fair comparison, we first conduct the experiment by only using images with known viewpoints as the input. Note that the work in [8] only incorporates the shoe images displayed in the 1^{st} and 4^{th} viewpoint (corresponding to the left-profile and left-facing viewpoint). We will carry a comparison of shoe images shown in these two views. For both of the systems, we extract the three complementary features (LBP, HOG and GIST) from the localized part for attribute prediction. The qualitative comparison results are shown in Table II. It can be seen that, compared with the work in [8], our proposed system achieves higher prediction accuracy for all the attributes when the inputs are with the 1^{st} and 4^{th} viewpoints. For the shoe image under other representative viewpoints, we can achieve over 90% attribute prediction accuracy for most of the attributes. Note that some of the shoe parts are not visible under certain representative viewpoints. For example, as to shoe images captured from the back view, the appearance of shoe head is hidden and not available. Therefore, the prediction of the corresponding attributes in those cases is not considered.

Next, we evaluate the performance of the proposed system without the knowledge of the viewpoints of the shoe images. The comparison results are shown in Table III. It can be seen that the attribute prediction accuracy degrades a bit comparing to the previous, which is due to the errors created from the viewpoint identification. However, the proposed system still works better than [8].

C. Discussions

The proposed system can be applied to tag the shoe images, which saves the merchants' effort in creating shoe descriptions. Fig.4 illustrate several examples of the shoes with the corresponding tags (i.e., the predicted attribute) using our approach. It can be seen that most of shoe tagging results are correct.

V. Conclusions

In this paper, we propose a novel semantic shoe tagging system capable of generating semantic attributes. We proposed to firstly identify its viewpoint and then predict the attribute values according to the identified viewpoint. During the attribute prediction, we propose a view-specific part localization model to localize the shoe part accommodating the attribute,

TABLE II
ATTRIBUTE PREDICTION ACCURACY(%) OF OUR PROPOSED SYSTEM AND [8] WITH KNOWN VIEWPOINT INPUT SHOE IMAGES

Attribute type	1^{st} View by [8] and Ours	4^{th} View by [8] and Ours	2^{nd} View by Ours	3^{rd} View by Ours	5^{th} View by Ours
Closed Toe	80.27**84.8**	83.21**86.8**	86.3	-	86.3
Back Cover	94.6**97.2**	95.7**97.6**	97.4	97.2	97.2
Side Cover	83.9**92.2**	82.4**90.1**	92.2	-	89.9
Pointy	87.2**90.2**	88.2**91.7**	91.4	-	92.0
High Thin Heel	80.2**89.7**	83.6**90.6**	-	88.33	90.4
Wedge Heel	86.7**94.7**	89.4**96.2**	-	92.5	96.4
Bounds	79.8**86.3**	80.6**86.9**	83.2	-	86.3

TABLE III
ATTRIBUTE PREDICTION ACCURACY(%) OF OUR PROPOSED SYSTEM AND [8] WITH UNKNOWN VIEWPOINT INPUT SHOE IMAGES

Attribute type	Unknown View by [8] and Ours
Closed Toe	78.3**85.2**
Back Cover	92.2**97.3**
Side Cover	83.2**91.4**
Pointy	86.8**89.9**
High Thin Heel	81.8**90.1**
Wedge Heel	84.2**95.2**
Bounds	76.2**86.1**

Fig. 4. Examples of shoe image tagging

from which complementary features are extracted and combined for the final attribute prediction. The evaluation on our newly constructed pump shoe dataset verifies the effectiveness of our approach.

ACKNOWLEDGMENT

This work was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by a grant from the Singapore National Research Foundation and administered by the Interactive & Digital Media Programme Office at the Media Development Authority.

REFERENCES

[1] Si Liu, Jiashi Feng, Zheng Song, Tianzhu Zhang, Hanqing Lu, Changsheng Xu, and Shuicheng Yan, "Hi, magic closet, tell me what to wear!," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 619–628.

[2] Basela Hasan and David Hogg, "Segmentation using deformable spatial priors with application to clothing.," in *BMVC*, 2010, pp. 1–11.

[3] Lukas Bossard, Matthias Dantone, Christian Leistner, Christian Wengert, Till Quack, and Luc Van Gool, "Apparel classification with style," in *Computer Vision–ACCV 2012*, pp. 321–335. Springer, 2013.

[4] Wei Di, Catherine Wah, Anurag Bhardwaj, Robinson Piramuthu, and Neel Sundaresan, "Style finder: Fine-grained clothing style detection and retrieval," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*. IEEE, 2013, pp. 8–13.

[5] Yan Wang, Sheng Li, and Alex C Kot, "Category-separating strategy for branded handbag recognition," in *Communications, Control and Signal Processing (ISCCSP), 2014 6th International Symposium on*. IEEE, 2014, pp. 61–64.

[6] Adriana Kovashka, Devi Parikh, and Kristen Grauman, "Whittlesearch: Image search with relative attribute feedback," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2973–2980.

[7] Adriana Kovashka and Kristen Grauman, "Attribute adaptation for personalized image search," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3432–3439.

[8] Junshi Huang, Si Liu, Junliang Xing, Tao Mei, and Shuicheng Yan, "Circle & search: Attribute-aware shoe retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 11, no. 1, pp. 3, 2014.

[9] Pedro Felzenszwalb, David McAllester, and Deva Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[10] Ross Girshick and Jitendra Malik, "Training deformable part models with decorrelated features," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3016–3023.