FASHION RECOMMENDATION ON STREET IMAGES

Huijing Zhan[1,*], *Boxin Shi*[2,*], *Jiawei Chen*[1], *Qian Zheng*[1], *Ling-Yu Duan*[2], *Alex C. Kot*[1]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2]National Engineering Lab for Video Technology, Peking University, Beijing, China
{hjzhan,jwchan,zhengqian,eackot}@ntu.edu.sg, {shiboxin,lingyu}@pku.edu.cn

ABSTRACT

Learning the compatibility relationship is of vital importance to a fashion recommendation system, while existing works achieve this merely on product images but not on street images in the complex daily life scenario. In this paper, we propose a novel fashion recommendation system: Given a query item of interest in the street scenario, the system can return the compatible items. More specifically, a two-stage curriculum learning scheme is developed to transfer the semantics from the product to street outfit images. We also propose a domain-specific missing item imputation method based on style and color similarity to handle the incomplete outfits. To support the training of deep recommendation model, we collect a large dataset with street outfit images. The experiments on the dataset demonstrate the advantages of the proposed method over the state-of-the-art approaches on both the street images and the product images.

Index Terms— Fashion Recommendation, Outfit Completion, Street Photos, Curriculum Learning

1. INTRODUCTION

Due to the huge profits brought by online shopping, visual research on fashion becomes increasingly popular. However, most of the studies focus on the attribute detection [1] and recognition [2], clothing parsing [3] and landmark localization [4], fashion retrieval [5], *etc*. Recently, there is an increasing demand for developing an automatic fashion recommendation system. Such fashion recommendation system brings benefits to both the customers and online merchants, since it not only provides clear targets for customers, but also encourages them to purchase an outfit.

The key task of a fashion recommendation system is to learn the compatibility relationships. Such compatibility relationship can be learned among items from different categories. Veit *et al.* [6] proposed to explore the cross-category relationships with the Siamese network [7]. McAuley *et al.* [8] developed a distance metric on the deep features extracted from Amazon co-purchase data to measure the compatibility. He *et al.* [9] explored the localized notions of “relatedness” for compatibility evaluation. To improve the representation power, Veit *et al.* [10] developed a shared embedding combined with mask learning to capture the compatibility in different fascades. Vasileva *et al.* [11] further proposed to learn a non-linear embedding space for each category pair with triplet networks. However, these approaches only allow the evaluation of the compatibility between every two categories of items, without considering the overall look as a whole. Han *et al.* [12] tackled this problem by exploiting the sequence of co-occurring product images rather than pairs of items. Hsiao *et al.* [13] proposed to generate the personal wardrobe from the fashion images in different scenarios.

However, most of the above methods mainly focus on the product images and ignore the street images captured in more complex and realistic daily life scenario . Fashion recommendation based on street images is quite challenging in three aspects: 1) The cluttered background distracts the recommendation model learning from clothing items; 2) The densely annotated meta-data are not available; 3) The training data contain many incomplete outfits.

In this paper, we propose a novel fashion recommendation system, which is capable of returning the compatible items given a query item of interest in the street scenario. We consider the street image as a composition of individual product item and the context information. To exclude the distractive effect of cluttered background, a fashion apparel detector is developed to detect and crop each product item. To address the problem of sparsely annotated outfits in the street images, we propose to develop a curriculum learning scheme to transfer the semantics from rich annotated product images to street images. And for the incomplete outfits, we propose a novel domain-specific missing imputation method based on color and style similarity. Moreover, we construct a dataset by collecting street images from the social media and verify our proposed framework on this dataset.

Our contributions are twofold: (1) We develop a fashion recommendation system using street images, which efficiently learns from images with cluttered background,

*Corresponding authors. This research was carried out at the Rapid-Rich Object Search (ROSE) Lab, supported by the National Research Foundation, Singapore, under its Interactive Digital Media (IDM) Strategic Research Programme. It is also supported in part by the National Natural Science Foundation of China under Grant 61872012 and 61661146005, and the NTU-PKU Joint Research Institute through the Ng Teng Fong Charitable Foundation.

978-1-5386-6249-6/19/$31.00 ©2019 IEEE

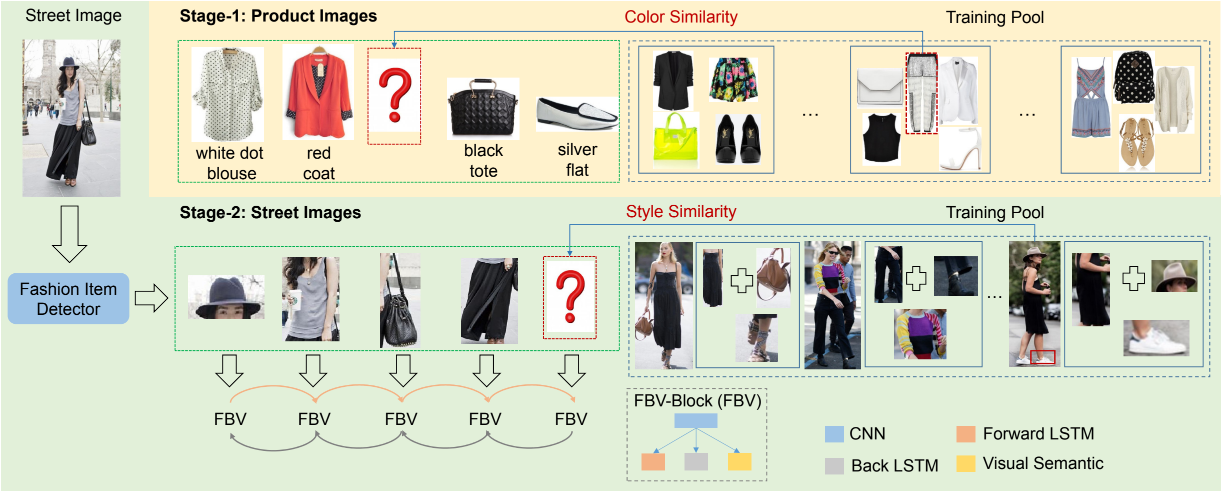

Fig. 1. The illustration for the pipeline of fashion recommendation system.

sparsely annotated and incomplete outfit items; (2) We build a large-scale dataset with street images in different occasions and regions.

2. PROPOSED METHOD

In this section, we will introduce a novel fashion recommendation system from the street images and the overall framework is illustrated in Fig. 1. The main function of this framework is to transfer the semantic knowledge learnt from the product images to the sparsely annotated street images via a two-stage curriculum learning strategy. Furthermore, to deal with the incomplete items (denoted as the red question mark in Fig. 1) in an outfit, we develop a domain-specific missing item imputation approach with style similarity on street images and color similarity on product images. The more details of each component will be introduced in the following paragraphs.

For a street outfit image I_i in the street scenario, a fashion item detector, trained with the pixel-level annotations, is utilized to detect and crop the individual items. Then it is represented as a sequence of m items as $I_i = [I_1^i, I_2^i, ..., I_m^i]$ in the pre-defined order (from top to bottom to accessories), which are subsequently fed into the Convolutional Neural Network (CNN) model for feature learning and grouped into a feature set, denoted as $\mathbf{X}_i = [\mathbf{x}_1^i, \mathbf{x}_2^i, \dots, \mathbf{x}_m^i]$. To model the compatibility relationships among these items, a bi-directional Long Short Term Memory (Bi-LSTM) network is introduced into the proposed framework. For the t th item $\mathbf{x}_t^i$, a forward LSTM predicts the next image given previous images and its loss function $F_1(\mathbf{X}_i;\Theta_1)$ can be expressed as below:

$$F_1(\mathbf{X}_i;\Theta_1) = -\frac{1}{m}\sum_{t=1}^{m}\log p(\mathbf{x}_{t+1}^i|\mathbf{x}_1^i, \mathbf{x}_2^i, \dots, \mathbf{x}_t^i;\Theta_1), \tag{1}$$

where Θ_1 denotes the parameters of the forward prediction model. Similarly, the backward LSTM predicts the backward images in the reverse order and the loss function can be written as below:

$$F_2(\mathbf{X}_i;\Theta_2) = -\frac{1}{m}\sum_{t=m-1}^{1}\log p(\mathbf{x}_t^i|\mathbf{x}_{t+1}^i, \mathbf{x}_{t+2}^i, \dots, \mathbf{x}_m^i;\Theta_2), \tag{2}$$

where Θ_2 indicates the parameters of the backward prediction model. Similar to [12], to exploit the multi-modal similarity (*e.g.,* images and the associated text descriptions), we define a joint visual-semantic embedding space as follows:

$$\begin{aligned} F_3(\Theta_3) = &\sum_u\sum_s \max(0, m - d(u,v) + d(u,v_s)) + \\ &\sum_v\sum_s \max(0, m - d(u,v) + d(u_s,v)), \end{aligned} \tag{3}$$

where Θ_3 means the parameters of visual-semantic model. $(u,\, v)$ is a paired image-text and v_s is the non-matching text with u. Similarly, u_s is the non-matching image to v. Thus, the fashion recommendation model can be trained by minimizing the sum of the above three loss functions denoted as below:

$$\min_{\Theta_1,\Theta_2,\Theta_3} F = \sum_i(\lambda_1 F_1(\mathbf{X}_i;\Theta_1) + \lambda_2 F_2(\mathbf{X}_i;\Theta_2)) + \lambda_3 F_3(\Theta_3), \tag{4}$$

where λ_1, λ_2, and λ_3 are the hyper-parameters that control the importance of forward, backward and visual-semantic losses, respectively, as shown in the FBV-Block (FBV) of Fig .1. In the proposed framework, we first learn the model parameters Θ_1, Θ_2 and Θ_3 on product images and then transfer them on street images.

Due to the ineffectiveness of the Faster-RCNN [14] detector on small objects [15] and clothes in large deformation [16], there exist many incomplete outfits. According to the prior knowledge of fashion matching rules, we evaluate whether an outfit is complete or not based on the two criterions: 1) Upper body (top, coat) + Lower body (short, skirt, pant) + Accessories (handbag, hat, shoes); 2) Dress + Accessories (handbag, hat, shoes). Note that at least one of the accessories is required. Due to the discrepancy in the presence of text descriptions (*e.g.,* meta-data), we deal with the incomplete street and product images in different manners described as below.

For the street images, due to the lack of the densely annotated meta-data, the missing items are completed based on the style similarity, which is measured by the StyleNet [17]. For notation simplicity, we denote the style extraction network as $f_s(\cdot)$. The overall style is evaluated by the composition of the individual items and the context information. Given a style descriptor $f_s(\cdot)$, the style similarity between a pair of street outfits I_i, I_j is calculated as below:

$$d(I_i, I_j) = \|f_s(I_i) - f_s(I_j)\|_2^2, \tag{5}$$

where $\|\cdot\|_2$ represents the ℓ_2 norm and smaller $d(I_i, I_j)$ indicates the more similarity between I_i and I_j.

For the product images, the incomplete outfits are completed based on the color similarity. The color similarity between a pair of items is categorized into three levels and each level will be assigned by a score $w_l, l = \{1, 2, 3\}$. Level 1 similarity indicates two colors are exactly the same. Level 2 similarity indicates that two colors are different but in compatible lists. That is, one of them is in black, white or grey color, which are suitable in most cases. Level 3 similarity indicates they are in different colors. And the color similarity score $F_c(i, j)$ between item i and item j is calculated as the score w_l with different level, denoted as below:

$$F_c(i,j) = \begin{cases} w_1, & \text{Level 1 Similarity} \\ w_2, & \text{Level 2 Similarity} \\ w_3, & \text{Level 3 Similarity}. \end{cases} \tag{6}$$

Therefore, the pairwise similarity between a pair of outfits I_i and I_j in the product images, denoted by C_{I_i,I_j}, is computed as below:

$$C_{I_i,I_j} = \sum_{i}^{m_i} \sum_{j}^{m_j} F_c(i,j), \tag{7}$$

where m_i and m_j denote the number of items in polyvore outfits I_i and I_j. And higher C_{I_i,I_j} value indicates the more similarity between the pairwise outfits I_i and I_j.

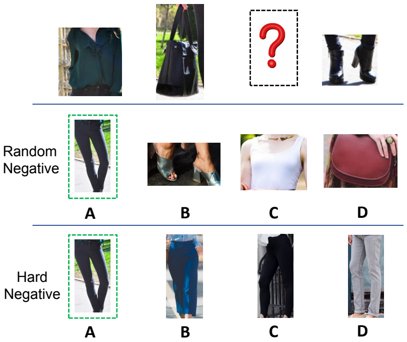

Fig. 2. Fill-in-the-blank evaluation metric. The matched ones are labeled in green bounding box.

In the test phase, given an item of interest in the street scenario, we first extract the embedding as the item representation, then it is fed forward into the Bi-LSTM to perform predictions in two directions. Finally, the output is a composition of items generated by forward and backward LSTM prediction.

3. EXPERIMENTS

3.1. Dataset and Experiment Settings

We collect the street outfit images from *Chictopia*[1] with occasion labels and *Wear*[2] with region-wise labels. Each item contains a product image and the outfit description. Finally, the dataset consists of 60,180 outfits in both asian (*e.g.,* China, Korea, Japan) and western (*e.g.,* UK and US) styles under the everyday scenario. For the product images, we utilised the Polyvore dataset [12] including 21,889 outfits with rich multi-modal information. After removing those images with irrelevant categories (*e.g.,* furniture, painting), 18,203 outfits are remained for training and evaluation.

Experiment Setup and Evaluation Protocol. Due to the lack of bounding box annotation on street images, we leverage the available fashion parsing dataset with pixel-level annotations as supervised information. The annotated training images are from three datasets: 1) Clothing Co-Parsing (CCP) Dataset [18] (1004 images); 2) Fashionista Dataset [19] (685 images); 3) Color-Fashion Dataset [20] (2682 images). Then they are further split into 4091 training and 280 testing images. The fashion apparel detector is built under the Faster-RCNN framework with the ResNet 50 as the backbone. The effectiveness of the apparel detector is based on the mean Average Precision (mAP).

For the Polyvore dataset, we conduct the experiments on 2,157 random negatives and 2,161 category-aware hard negatives. For the street dataset, 2,100 random negatives and

[1] http://www.chictopia.com
[2] http://www.wear.jp

Table 1. Comparisons with state-of-the-art method for Random Negative (RN) and Hard Negative (HN) Examples.

Approach	FITB Accuracy		AUC Score	
	RN	HN	RN	HN
Bi-LSTM [12]	0.73	0.68	0.80	0.84
CSN [11]	0.69	0.70	**0.94**	**0.92**
Proposed	**0.77**	**0.72**	0.83	0.86

Table 2. Evaluating the effect of the Curriculum Learning (CL) and the Missing Item Imputation (MII) on the street (the first three rows) and Polyvore dataset (the last two rows).

Dataset	Approach	FITB Accuracy		AUC Score	
		RN	HN	RN	HN
Street	Proposed w/o MII	0.74	0.69	0.81	0.84
	Proposed w/o CL	0.75	0.70	0.82	0.86
	Proposed	**0.77**	**0.72**	**0.83**	**0.86**
Polyvore	Proposed w/o MII	0.73	0.44	0.93	0.94
	Proposed	**0.78**	**0.46**	**0.95**	**0.96**

2,100 category-aware hard negatives are chosen for experiments. In the random negative examples, no prior knowledge of the missing blank (category) is provided; The hard negative indicates that the category of the missing item is known, the potential candidates can only be chosen in the same category. For the Polyvore dataset, the recommendation is pre-trained on the InceptionV3 model. For the hyper-parameters of Bi-LSTM model, we follow [12] as the guideline.

For quantitative evaluation, two metrics are utilized to estimate the performance of the fashion recommendation system, 1) FITB (fill-in-the-blank) accuracy, as shown in Fig. 2. Given multiple query items in an outfit (one item is erased, illustrated as the question mark), the system is expected to choose the right answer out of 4 choices; 2) fashion compatibility prediction accuracy, which is evaluated by the area under the ROC curve (AUC).

3.2. Experimental Results

For the fashion apparel detector, we set the correct detection Intersection over Union (IoU) threshold to 0.5. Our detector achieves 74.8% mAP score on the test images. For the fashion recommendation model, we compare the performance of the proposed method with two state-of-the-art algorithms: 1) sequence-based Bi-LSTM network [12] and 2) metric learning based CSN network [11].

We can find that the experiments conducted on the street dataset demonstrate that our proposed method achieved better performances in terms of FTIB accuracy with the other two state-of-the-art methods and comparable AUC scores.

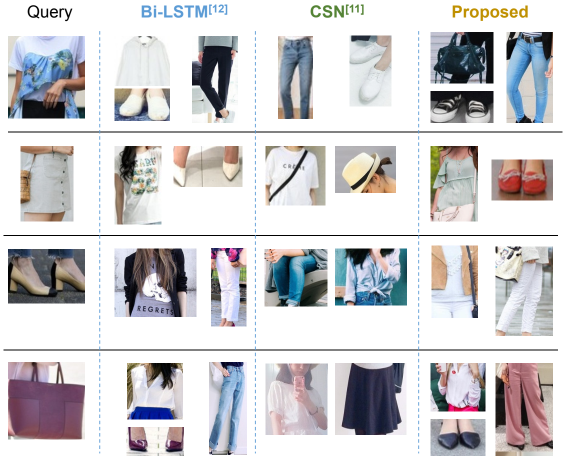

Fig. 3. Fashion recommendation examples given the query item from the street scenario. Best view in color.

Fig. 3 demonstrate the recommended outfits with the query items in the street scenario. It can be seen that the proposed system is capable of recommending the outfits in variable styles and the overall styles of different fashion items in an outfit match with each other. Compared with the other two approaches, our method puts an emphasise on the overall style of the outfits rather than merely the color similarity.

3.3. Effectiveness of the Curriculum Learning and Missing Item Imputation

We evaluated the effectiveness of the curriculum learning and missing item imputation component in the model learning, as shown in the first three rows of Table. 2. It verifies the effectiveness of transferring semantics from the product images with densely annotated meta-data. Also, the imputation strategy based on the style similarity also improves the performance of recommendation. The results on the Polyvore dataset in the last two rows of Table 2 further verify the effectiveness of the imputation strategy on the product images.

4. CONCLUSION

In this paper, we propose a novel fashion recommendation system to return the compatible items given a query item in the street scenario. We transfer the high-level semantics from product images to sparsely annotated street images via the two-stage curriculum learning strategy. To deal with the incomplete outfits, we employ the style and color similarity as the guideline to perform the missing item imputation. Moreover, we build a large-scale dataset with street images in different occasions and region. The experimental results demonstrate the advantages over the state-of-the-art approaches.

5. REFERENCES

[1] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis, "Automatic spatially-aware fashion concept discovery," in *Proceedings of International Conference on Computer Vision*, 2017.

[2] Qi Dong, Shaogang Gong, and Xiatian Zhu, "Multi-task curriculum transfer deep learning of clothing attributes," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2017.

[3] Si Liu, Xiaodan Liang, Luoqi Liu, Ke Lu, Liang Lin, Xiaochun Cao, and Shuicheng Yan, "Fashion parsing with video context," *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1347–1358, 2015.

[4] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[5] Huijing Zhan, Boxin Shi, and Alex C Kot, "Cross-domain shoe retrieval with a semantic hierarchy of attribute classification network," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5867–5881, 2017.

[6] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie, "Learning visual clothing style with heterogeneous dyadic cooccurrences," in *Proceedings of International Conference on Computer Vision*, 2015, pp. 4642–4650.

[7] Sean Bell and Kavita Bala, "Learning visual similarity for product design with convolutional neural networks," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, pp. 98, 2015.

[8] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel, "Image-based recommendations on styles and substitutes," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015.

[9] Ruining He, Charles Packer, and Julian McAuley, "Learning compatibility across categories for heterogeneous item recommendation," in *Data Mining, IEEE 16th International Conference on*, 2016.

[10] Andreas Veit, Serge Belongie, and Theofanis Karaletsos, "Conditional similarity networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[11] Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth, "Learning type-aware embeddings for fashion compatibility," *arXiv preprint arXiv:1803.09196*, 2018.

[12] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis, "Learning fashion compatibility with bidirectional lstms," in *Proceedings of the ACM on Multimedia Conference*, 2017.

[13] Wei-Lin Hsiao and Kristen Grauman, "Creating capsule wardrobes from fashion images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[15] Shuai Zheng, Fan Yang, M Hadi Kiapour, and Robinson Piramuthu, "Modanet: A large-scale street fashion dataset with polygon annotations," *arXiv preprint arXiv:1807.01394*, 2018.

[16] Kota Hara, Vignesh Jagadeesh, and Robinson Piramuthu, "Fashion apparel detection: the role of deep convolutional neural network and pose-dependent priors," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2016.

[17] Edgar Simo-Serra and Hiroshi Ishikawa, "Fashion style in 128 floats: joint ranking and classification using weak data for feature extraction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[18] Wei Yang, Ping Luo, and Liang Lin, "Clothing co-parsing by joint image segmentation and labeling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3182–3189.

[19] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg, "Parsing clothing in fashion photographs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[20] Si Liu, Jiashi Feng, Csaba Domokos, Hui Xu, Junshi Huang, Zhenzhen Hu, and Shuicheng Yan, "Fashion parsing with weak color-category labels," *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 253–265, 2014.