1.0 inch

Visual Product Search for Fashion Domain: From Hand-crafted features to Deep Learning

Huijing Zhan†, Yan Wang†, Abrar H. Abdulnabi†, Sheng Li*, Dennis Sng*, Alex C. Kot* and Simon See#

†Rapid-Rich Object Search (ROSE) Lab
Nanyang Technological University
Singapore
{Zh0069NG, WANG0696, ABRARHAM001}@e.ntu.edu.sg

*School of EEE
Nanyang Technological University
Singapore
{EACKOT, DENNIS.SNG, LISHENG}@ntu.edu.sg

#NVIDIA Corporation, Santa Clara, USA
ssee@nvidia.com

0.2 inch

***Abstract*— Due to the huge profits from e-commerce especially in the fashion domain, a large body of research has been conducted to enable computers to intelligently perceive and analyze the objects visually, which involves multi-disciplines like computer science and artificial intelligence. Also, different categories of fashion products have been explored like clothing, handbag and shoes. For efficient description of those objects, the selection of an appropriate feature representation scheme is of great importance. In recent years, due to the unmatched performance of deep learning methods, the trends move from using the traditional hand-crafted features to automatically exploring the appropriate feature for representation by analyzing the big data. In this paper, we use inter-disciplinary knowledge from computer science, engineering and system architecture to study fashion-related object search. Existing techniques are reviewed and new methods are introduced. Preliminary experimental results on product search are presented using both traditional and deep learning approaches.**

***Keywords*— visual object search, e-commerce, deep learning, computer vision, fashion search**

I. INTRODUCTION

1.0 inch

Fashion profits have occupied a large portion of the entire market. It is reported that in the clothing market alone, total retail sales have reached up to about $980 billion in 2012 [1]. Due to its promising market size, research in the fashion field has been receiving more and more attention, especially the intelligent fashion analysis. It enables computers to analyze the fashion product more smartly using vision-based approaches, which incorporate multi-discipline knowledge from computer science, engineering. Currently, researchers in this field have been developing algorithms addressing different types of fashion products, like handbags [6], shoes [9], clothing [2], [3] and the list will go on.

To efficiently employ vision based methods on objects, firstly it is essential to develop an appropriate feature representation scheme. Conventionally, hand-crafted features like SIFT, HOG, GIST are popular choice for their powerful ability to capture

the salient semantics of objects. However, in recent years, deep learning based methods [10], which model high-level abstractions of objects [11], have become increasingly popular, outperforming state-of-the-art traditional methods in handling various vision tasks and different objects as noted in [13], [14]. What is more, it makes good use of multi-disciplinary knowledge like neural networks in computer science, borrowing biological concepts to model how humans perceive objects, and system architecture from engineering aspects.

Due to the significant performance of deep learning based methods and the promising potential in the fashion industry, the Rapid-Rich Object Search (ROSE) Lab has dedicated resources to address fashion related product search problems, like handbag recognition, clothing retrieval, shoe tagging, *etc*.

II. Hand-Crafted Features vs Deep Learning Features

Traditional pipeline for the fashion product search uses the hand-crafted features like HOG, SIFT to represent the characteristics of the objects. Even though they have demonstrated their powerfulness of representation capability on a variety of visual search tasks, they are still unable to carry high-level concepts of objects. To address this problem, in recent years, researchers proposed a neural network based deep learning architecture named Convolutional Neural Networks (CNNs).

The architecture [10] of classical deep learning based approaches contains eight layers inclusive of five convolutional and three fully-connected layers. The left-most is an input image, and after traversing through all the eight layers, the network predicts it into one of the 1000 categories according to the ILSVRC settings. We adapt the number of categories according to the specific visual search tasks we are faced with. In this way, the architecture could be well fitted to the different fashion product visual search tasks.

III. Deep Learning Resources

There are many existing tools which can facilitate GPU accelerated deep learning, including Caffe [6], Cuda-convnet [12], MatConvNet [7] and Nvidia Digits [8]. Caffe is a c++ deep learning framework capable of custom network definition using Google protocol buffers. Cuda-convnet is a c++ - based project that allows the definition of deep neural network architectures using configuration files. MatConvNet is an implementation of Convolutional Neural Networks (CNNs) for MATLAB, which is designed with an emphasis on simplicity and extendibility. Nvidia Digits is a software based on Caffe, which provides an interface for data preparation, network configuration and visualization of training process. These tools have their own pros and cons. All of them allow GPU accelerated deep learning, which is much faster than using CPUs alone. Cuda-convnet and caffe have a relatively higher speed compared with MatConvNet, while Matconvnet is the most simple and easy tool for customized deep learning.

Nvidia's next generation GPUs have significant architectural improvements to facilitate deep learning. Nvidia's next GPU architecture Pascal will support up to 32 GB of high bandwidth memory. With an increase in memory size, support for FP16 instructions and higher memory bandwidth, there will be significant performance improvements in deep leaning models in future Nvidia GPUs. This performance will increase further by splitting work across multiple GPUs and having a dense system where GPUs are connected by NVLINK, Nvidia's new point-to-point interconnect. NVLink will enable very dense compute farms where a single node can have 8 or more GPUs.

IV. Fashion Product Search and Results

In this paper, we present the research at the ROSE Lab on visual search for fashion related products, specifically, on handbag, shoes and clothes. We will also demonstrate their correlated applications and demos for illustration.

A. Handbag

For those companies well-known for selling luxury handbags, one of the main issues is how to make their handbags more popular to their customers. This will require the manufacturers to gain a deep insight into the customers' feedback about their handbags. A traditional way to get the consumers' comments about one particular handbag is to do a keyword based search through the internet. However, nowadays people prefer to upload the photos of their purchases on blogs or twitters without describing the specific names or models, which brings difficulties for manufacturers to collect users' feedback. Therefore, it would be necessary to develop an image based handbag recognition engine for these manufacturers.

In order to identify the subtle differences among handbags, we propose to enhance the handbag local structure pattern, and extract the feature from the enhanced handbag image to complement the feature extracted directly from the original handbag image. We term these two types of features as the complementary and original features. We also test different deep learning features for handbag recognition.

1) Branded Handbag Dataset: We build a branded handbag dataset for one particular brand: Louis Vuitton. Images in our dataset are downloaded from Google, Flicker as well as some online shopping websites. Eventually, the dataset contains 220 handbag models (classes). We randomly partition our handbag dataset into two parts, five images per handbag for training and the rest for testing. Fig.1 shows the examples handbag images in the database.

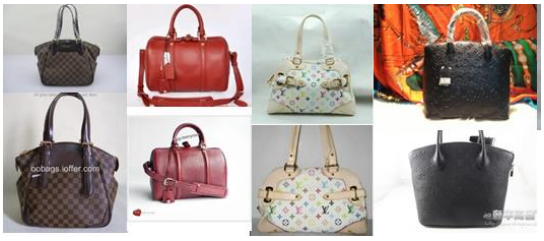

Fig.1. Examples of handbag images in the database

2) Experimental Results: The performance of the branded handbag recognition using each single feature and linear SVM classifier are given in Table I. It can be seen that using the complementary feature alone will get comparable accuracy with using the original feature. The concatenation of the original feature and the complementary feature (e.g., SIFT & Complementary SIFT) performs consistently better than the single original feature, with over 2.5% improvement in accuracy. We also summarize the recognition accuracies of the 6th layer and 7th layer Decaf features. The 6th layer feature leads to better results (around 6% better) than the 7th layer feature, and is superior to traditional features.

TABLE I

Handbag recognition accuracies (%) using traditional and deep learning features.

Features	**Accuracy**
SIFT	70.02
Complementary SIFT	67.43
SIFT & Complementary SIFT	72.77
HOG	63.99
Complementary HOG	60.68
HOG & Complementary HOG	67.45
Decaf6	82.70
Decaf7	76.94

3) Related Applications: A user could take a photo of a handbag using a mobile phone and find where to buy this handbag. Fig. 2 demonstrates the handbag recognition flowchart.

Fig.2. Mobile-based handbag recognition demo capture

B. Shoes

The development of efficient shoe tagging techniques is not only important to buyers to enrich the shopping experience, but also for the online merchants selling fashion items, where a great many new products are uploaded every day. Here, the annotation of shoes corresponds to providing labels for a

1.0 inch

shoe, for example, given a pump shoe input image, the computer is capable of automatically generating descriptions like `high heel' and'has back cover'. Few works focus on addressing the shoe annotation tasks, which describe high-level concepts of shoes.

We build a novel a multi-view shoe dataset, on which traditional and deep learning features are tested to evaluate the attribute prediction result.

1) Pump Shoe Dataset: The images of our pump shoe dataset are collected from Amazon.com with clean background. In total, our pump shoe dataset consists of 7500 shoe images in multiple viewpoints. For each shoe, the groundtruth annotation contains seven binary part-aware attributes defined based on four different shoe parts, as shown in Table II. The groundtruth annotation of each attribute is collected manually. Fig.3 displays the multiple viewpoints settings for a shoe.

Fig.3. Multiple-view display settings for a shoe

TABLE II
Shoe parts and their part-awareness attributes

Shoe Parts	Related Attributes
head	Closed toe, pointy
body	Side-covered, bounds
Back	Back-covered
heel	High thin heel, wedge heel

2) Experimental Results: Table III shows the results of shoe attribute annotation using the traditional features SIFT and HOG in comparison with deep learning features $Decaf_7$ and $Decaf_8$ activated from the full-connected layers fc7 and fc8 of the deep learning multi-layer networks. The performance is evaluated using the Mean Average Precision metrics. In the Table, we use the bold numbers to highlight the best-performing feature. From the results, we could find that even though traditional features performs well on annotation tasks, however, employing deep learning features improved attribute prediction significantly. Also, among all the attributes defined, $Decaf_8$ achieves slightly higher prediction accuracy than $Decaf_7$. It is reasonable for the last full-connected layer to provide more powerful and hierarchical representations for shoe shapes.

TABLE III
Attribute prediction accuracy (%) using traditional and deep learning features

Attribute Type	HOG	SIFT	Decaf7	Decaf8
Closed toe	83.91	80.36	86.23	**87.13**
Pointy	91.18	90.61	90.73	**91.20**
Bounds	87.87	87.17	88.90	**89.22**
High-heel	88.50	88.32	**90.52**	90.37
Wedge-heel	94.95	94.37	**95.83**	94.78
Side-covered	89.98	88.90	94.10	**94.48**
Back-covered	95.34	94.93	96.64	**97.97**

3) Related Applications: The proposed system can be applied to tag the shoe images, which saves the merchants' effort in creating shoe descriptions.

C. Clothing

In on-line fashion shopping, clothes such as dresses are top selling products among feminine market place. This leads to the need to develop vision programs that are able to predict which dress items are more attractive, thus having the potential to be placed on the top of the website pages.

Fig.4. Example of clothed people images from the Clothing Attribute database [5]

1) Experimental Result: To tackle the problem of extracting robust attribute feature representations, we adapt deep feature learning methods like convolutional neural

networks (CNN) [10]. Specifically, we adapt deep CNN to learn binary semantic attributes, where each CNN will predict one binary attribute, hence each CNN will generate attribute-specific feature representations. We train these CNN models on the Clothing attribute dataset [5] including clothed people mostly pedestrians on the street with cluttered background as shown in Fig.4. In table IV, we show the mean average accuracy prediction of attributes to compare the performance of some baseline method, CNN models and a state-of-the-art method. S-CNN refers to trained attribute-specific models of CNN. CF refers to the combined features model with no pose baseline [4], while CRF refers to the previous state-of-the-art method proposed by [5].

TABLE IV
The mean average accuracy (%) of attribute prediction on the Clothing dataset [23].

Method	**Accuracy**
S-CNN [10]	90.43
CF [4]	80.48
CRF [5]	83.95

2) Related Applications: A direct application that can leverage the learned semantic attribute representations, is to train a ranking algorithm to predict which dress images are more likable than others on our collected datasets.

ACKNOWLEDGMENT

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, Prime Minister's Office, Singapore, under its IDM Futures Funding Initiative and administered by the Interactive and Digital Media Programme Office.

We also gratefully acknowledge the support of NVIDIA Corporation for their donation of Tesla K40 GPUs used for our research at the ROSE Lab.

REFERENCES

[1] Si Liu, Luoqi Liu, and Shuicheng Yan, "Fashionanalysis: Current techniques and future directions," *MultiMedia*, IEEE, vol. 21, no. 2, pp. 72–79, 2014.

[2] 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.

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

[4] A. Gallagher and T. Chen. "Clothing segmentation for recognizing people", in *Computer Vision and Pattern Recognition, 2008.CVPR 2008. IEEE Conference on.* IEEE, 2008, pp. 1–8.

[5] H. Chen, A. Gallagher, and B. Girod. "Describing clothing by semantic attributes,"in *Computer Vision–ECCV 2012*, pp.609–623.Springer, 2012.

[6] http://caffe.berkeleyvision.org/

[7] http://www.vlfeat.org/matconvnet/

[8] https://github.com/NVIDIA/DIGITS

[9] 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.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks". In *NIPS*, pages 1097–1105, 2012.

[11] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. "Cnn features off-the-shelf: an astounding baseline for recognition". arXiv preprint arXiv:1403.6382, 2014.

[12] https://code.google.com/p/cuda-convnet/

[13] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. "Deepface: Closing the gap to human-level performance in face verification". In *CVPR*, pages 1701–1708, 2014.

[14] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. "Panda: Pose aligned networks for deep attribute modeling". In *CVPR*, 2014.