COMPLEMENTARY FEATURE EXTRACTION FOR BRANDED HANDBAG RECOGNITION

Yan Wang, Sheng Li, and Alex C. Kot

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

ABSTRACT

Fine-grained object recognition aims at recognizing objects belonging to the same basic-level class such as dog, bird or fish, which is a challenging problem in computer vision. In this paper, we consider the problem of recognizing handbags that belong to a specific brand. In order to identify the subtle differences among handbags, we propose to enhance the handbag local structure pattern by using the Hölder exponent, and extract the feature from the enhanced handbag image to complement the feature extracted directly from the original handbag image. We term such two types of features as the complementary and original features. These features will then be fused by using Multiple Kernel Learning (MKL) for branded handbag recognition. We conduct the experiments on a newly built branded handbag dataset, the results of which demonstrate the effectiveness of the proposed complementary feature in recognizing the handbags.

Index Terms— Handbag, Fine-grained Object Recognition, Hölder Exponent, MKL

1. INTRODUCTION

For those companies well-known of selling luxury handbags, one of the main issues is how to make their handbags more popular and acceptable 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.

Generally speaking, there are two kinds of object recognition techniques, i.e., generic object recognition and fine-grained object recognition, which differ in handling different levels of object categories. Generic object recognition aims to recognize gross differences of distinct object categories [1, 2]. While fine-grained object recognition concentrates on dealing with subtle differences among highly similar object classes [3, 4, 5], which poses even more challenges to human beings.

Fig. 1. Examples of handbags with different patterns. Handbags with (a) Checkerboard pattern, (b) Embossed horizontal texture pattern and (c) Quatrefoils and flowers.

The aforementioned branded handbag recognition falls into the category of fine-grained object recognition. In recent years, more and more efforts have been devoted on the fine-grained object recognition problem. Researchers propose various techniques for recognizing the objects within one basic category such as bird [4], food [6], or dog [5]. These techniques try to identify the subtle differences among different sub-categories, where the use of effective features to represent object patches or parts is critical for the recognition. Since for handbags, the subtle texture difference among different handbags is non-trivial for the recognition, extracting proper features to describe local structure patterns tends to be the key to significantly improve the recognition performance.

Various features have been proposed for recognition tasks, including SIFT [7], HOG [8] and LBP [9]. SIFT measures

This research is supported by the Singapore National Research Foundation under its IDM Futures Funding Initiative and administered by the Interactive & Digital Media Programme Office, Media Development Authority.

the appearance of a point by computing the weighted spatial histogram of gradients in its neighborhood. It is invariant to the rotation and illumination. HOG captures the intensity gradient structure and edge direction of the local shape with uniform sampling and fine orientation binning. Orientation based features like SIFT and HOG are powerful because they are robust to changes in brightness. LBP is used to analyze the texture which labels the pixel through utilizing the gray scale contrast of the neighborhood and this pixel. It thresholds the neighboring pixel and saves the result as a binary number [10]. It is believed to perform well under some monotonic gray-scale changes. Some other features like rotationally invariant Maximum Response filter sets (Texton) are also proved to perform well for discriminating isotropic as well as anisotropic textures [11].

However, these features only capture the local region information and may not pay enough attention to the details of the textured regions in images. In addition, one single feature is not sufficient for describing the structures as well as the detailed patterns of objects, fusing multiple complementary features would be useful for fine-grained object recognition. In this work, we try to tackle the problem of branded handbag recognition, which is rarely studied in the literature. To identify the subtle differences among different handbags of the same brand, we propose to extract the feature from the structure enhanced handbag image. The structure enhanced handbag image is obtained by applying Hölder exponent to the original image. This feature, called complementary feature, is used to complement the details of the pattern that are not captured by the feature, called original feature, extracted directly from the original handbag image. Complementary feature is then fused with the original feature by using Multiple Kernel Learning (MKL) for branded handbag recognition. We evaluate the effectiveness of the proposed complementary feature for several popular features on a newly constructed handbag dataset. The results show that fusing the complementary and original features could always boost the accuracy of the branded handbag recognition.

2. COMPLEMENTARY FEATURE EXTRACTION

The variation of patterns on handbags is large. Take the brand [12] considered in this paper as an example, there are handbags with checkerboard pattern, embossed horizontal texture pattern, quatrefoils and flower pattern as shown in Fig. 1. However, some of the patterns are not presented prominently. Embossed horizontal texture pattern is difficult to be identified because those embossed horizontal lines are subtle. Furthermore, as handbag photos are always taken under uncontrolled environment, the details of the patterns maybe decayed due to the lighting condition or out of focus.

In this section, we propose a complementary feature which is extracted from the structure enhanced handbag image, so as to cater the need of capturing the details of different patterns and facilitate the recognition. We propose to enhance a handbag image by using Hölder exponent from the multifractal theory [13]. This feature enhances the local structure by measuring the neighborhood gradient of a point based on the spatial interactions between the pixel and its neighboring pixels. The Hölder exponent of a grayscale image describes the regularity of the local structure in the image, which is computed as follows.

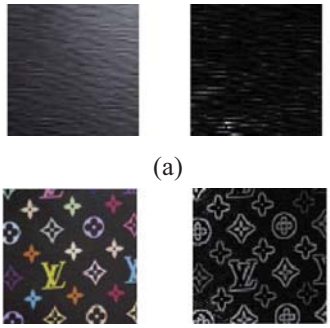

Fig. 2. Local structure enhancement by using the Höler exponent. Left: image patch, right: α-image. (a) Embossed horizontal texture pattern, (b) Quatrefoils and flowers pattern.

$$\alpha = \lim_{\varepsilon \to 0} \frac{\ln\left(\hbar(S_n)\right)}{\ln\left(\varepsilon\right)}. \tag{1}$$

where S_n refers to a local squared neighborhood with the size ε, $\hbar(S_n)$ is a local regularity measure for S_n. In our implementation, $\hbar(S_n)$ computes the maximum pixel value within S_n, i.e.,

$$\hbar(S_n) = \max_{(k,l)\in S_n} g(k,l) \tag{2}$$

where $g(k,l)$ is the pixel value located at (k,l). As indicated in Stojic's work [13], Hölder exponent has the ability to enhance the image local structure. It works well even when the differences among local neighborhood are subtle.

The Hölder exponent of the intensity of an image is called an α-image in multifractal theory, which appears in the form of a graycale image. Fig.2 illustrates two local image patches (with different patterns) and their corresponding α-images. It can be seen that the α-image effectively enhances the local structure pattern, where the embossed horizontal texture pattern displays as the horizontal arrangements and the quatrefoils and flower pattern appears more distinctively. After the structure enhancement, the complementary feature can be extracted from the α-image of the handbag using any existing feature extractor. In this paper, we choose several popular features including SIFT [7], HOG [8], LBP [9] and Texton [11]. For simplicity, we term the corresponding complementary feature as Hölder-SIFT, Hölder-HOG, Hölder-LBP and Hölder-Texton, respectively. While the features extracted

Fig. 3. Top 15 salient windows for handbag images (displayed in yellow bounding boxes) by using the method proposed in [14].

from the original handbag images are defined as the original features, which are termed as SIFT, HOG, LBP and Texton for short.

It should be noted that, instead of applying feature extraction methods densely from the whole handbag image (or the corresponding α-image), we adopt a saliency detection technique [14] and extract the features only from the salient region of the handbag image. Feng *et al.* [14] use a principled sliding window based paradigm, which detects salient objects as the windows corresponding to the local maxima. Examples of handbag saliency detection results using this method are given in Fig. 3. We experimentally choose only the top 15 saliency windows for each handbag image. The final salient region is the bounding box covering all the 15 windows.

3. FEATURE LEVEL FUSION

Once we extract the complementary feature, we will next fuse it with the corresponding original feature by using a feature level fusion algorithm. In recent years, multiple kernel learning (MKL) [15, 16] is very popular for the feature level fusion. Given several predefined base kernels, MKL is to train the SVM classifier and the kernel combination coefficients simultaneously. The soft margin MKL introduces a slack variable for each of the base kernels, allowing some errors for the training data, which makes it more robust in real applications. Therefore, we here adopt the state-of-the-art soft margin MKL method proposed by Xu *et al.* [16] for the feature level fusion, where the objective function is

$$\begin{aligned} &\min_{\tau,\boldsymbol{\alpha}\in A,\zeta_m} -\tau + \theta \sum_{m=1}^{M} \zeta_m \\ &s.t. -\frac{1}{2}(\boldsymbol{\alpha}\odot\boldsymbol{y})'\boldsymbol{K}_m(\boldsymbol{\alpha}\odot\boldsymbol{y}) \geq \tau - \zeta_m, \zeta_m \geq 0, m = 1, ..., M \end{aligned} \tag{3}$$

where τ denotes the target margin, $\boldsymbol{\alpha}$ is the coefficient associated with training samples, ζ_m is the slack variables for different kernels $\boldsymbol{K_m}$, and θ balances the loss term. Please refer to [16] for more details.

4. EXPERIMENTAL RESULTS

4.1. Dataset Construction

We newly construct a dataset for one particular brand [12]. We observe that there are totally 551 women handbags with different models displayed on the official website of this brand, where some of the handbags share the same style name. The appearance of the handbags belonging to the same style are similar with each other, differing mainly by the size or the color. We randomly pick one handbag out of each style, which serves as a representative handbag for this style. In total, there are 80 representative handbags covering all the existing styles. We construct a dataset consisting of 976 images for these representative handbags, where the images are downloaded online (from Google, Flicker, etc). The number of images of each handbag differs according to the online source. Handbag images in our dataset are mainly frontal view with slight rotation because we believe that this is the most important view when advertising a handbag. It allows us to focus on the pattern of handbags for handbag recognition. In the future, we are going to build a larger handbag dataset covering all the existing models displayed on the website [12].

4.2. Experimental settings

We randomly partition our handbag dataset into two parts: three images per handbag for training and the rest for testing. To be consistent with the previous work [17], we use the mean-average precisions (mAPs) to evaluate the performance.

After the salient detection, we normalize the image into the size of 256×256. Next, we built a 300-word dictionary, followed by adopting locality-constrained linear coding and three level spatial pyramid with max-pooling to learn the bag-of-feature descriptors for all local features. It should be noted that the current spatial pyramid technique [2] employs a decomposition strategy which partitions the feature space into 1×1, 2×2 and 4×4 cells. However, in our case, we exploit a decomposition approach with 1×1, 2×2, and 3×3 cell partitioning strategy because we find that this is the best spatial pyramid strategy for handbags.

For feature level fusion, we randomly choose the parameter θ within its domain for the soft margin MKL [16]. We employ one-vs-all SVMs for training and recognition the handbags. Histogram intersection kernel [18] is adopted as the base kernel for each of the features to be fused.

4.3. Performance evaluation of different features

The performances of the branded handbag recognition using each single feature (original or the complementary) are given in Table 1. It can be seen that using the complementary feature alone will get comparable accuracy with using the original feature. Table 2 shows the results by fusing the origi-

Table 1. Using each single feature for branded handbag recognition

Feature	SIFT	HOG	LBP	Texton
mAP (%)	**85.93**	76.06	44.83	38.49
Feature	Hölder-SIFT	Hölder-HOG	Hölder-LBP	Hölder-Texton
mAP (%)	**83.18**	71.31	49.34	66.71

Table 2. Fusing the original feature and its complementary feature for branded handbag recognition

Feature fusion	mAP(%)
SIFT & Hölder-SIFT	**88.22**
HOG & Hölder-HOG	77.73
LBP & Hölder-LBP	49.76
Texton & Hölder-Texton	67.13

Table 3. Fusing the SIFT with other original features for branded handbag recognition

Feature fusion	mAP(%)
SIFT & HOG	85.91
SIFT & LBP	80.96
SIFT & Texton	85.93
SIFT & HOG & LBP	79.94
SIFT & HOG & Texton	85.91
SIFT & LBP & Texton	80.96
SIFT & HOG & LBP & Texton	79.94

nal feature and its complementary feature. By comparing Table 1 and Table 2, we can see that using such a fusion always achieves a higher mAP than using a single feature, which illustrates the good complementarity of our proposed complementary feature for the original feature. We find that SIFT based features (SIFT, Hölder-SIFT and SIFT & Hölder-SIFT) achieve the best performance when compared with other single or fused (original & complementary) features. Next, we fuse the SIFT with other original features, the results of which are given in Table 3. It can be seen that fusing other original features with SIFT is not helpful for the branded handbag recognition. This further demonstrates the effectiveness of our complementary feature for SIFT.

4.4. Applying complementary feature on the existing fine-grained recognition system for branded handbag recognition

In this section, we apply our complementary feature on a popular and public available fine-grained recognition system [17]. This work proposes a general framework for fine-grained object recognition and gets PASCAL Challenge Winning Prize in the year of 2011 as well as in 2012. It applies randomization to sample a subset of image patches by designing random forests with strong classifiers, and find image regions that contain discriminative information for fine-grained object recognition. SIFT is chosen as the base feature for recognition in their paper. We apply this method directly on the salient region of each handbag image using the SIFT and SIFT & Hölder-SIFT for the recognition. It can be observed from Table 4 that SIFT & Hölder-SIFT performs consistently better than the SIFT feature regardless of the tree numbers.

Table 4. Applying complementary feature on the existing fine-grained recognition system [17] for branded handbag recognition

# of decision trees	mAP (%)	
	SIFT	SIFT & Hölder-SIFT
95	81.99	85.15
96	82.08	85.15
97	82.25	85.26
98	82.30	85.00
99	82.53	84.94
100	82.48	85.27

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have worked on the branded handbag recognition, which is a fine-grained object recognition problem rarely investigated in the literature. To capture the detailed structures of the handbag, we have proposed a complementary feature extracted from the structure enhanced handbag image. Such feature is able to compensate the missing structure details of the original feature extracted from the original handbag image. Experimental results on a newly built branded handbag dataset demonstrate that, by fusing the proposed complementary feature and the original feature, we can achieve a better performance than using a single feature for branded handbag recognition. In the future, we will expand our dataset to cover more handbags. We will also investigate the effectiveness of the propose complementary feature for other fine-grained object recognition problems.

6. REFERENCES

[1] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," in *ICCV'13: Proc. IEEE 14th International Conf. on Computer Vision*, December 2013.

[2] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, vol. 2, pp. 2169–2178.

[3] B. Yao, G. Bradski, and L. Fei-Fei, "A codebook-free and annotation-free approach for fine-grained image categorization," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 3466–3473.

[4] R. Farrell, O. Oza, N. Zhang, V.I. Morariu, T. Darrell, and L.S. Davis, "Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, pp. 161–168.

[5] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 3498–3505.

[6] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 2249–2256.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition (CVPR) 2005 IEEE Conference on*, 2005, vol. 1, pp. 886–893 vol. 1.

[9] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.

[10] W. Shen, B. Wang, Y. Wang, X. Bai, and L. J. Latecki, "Face identification using reference-based features with message passing model," *Neurocomputing*, vol. 99, pp. 339–346, 2013.

[11] J.-M. Geusebroek, A. W M Smeulders, and J. van de Weijer, "Fast anisotropic gauss filtering," *Image Processing, IEEE Transactions on*, vol. 12, no. 8, pp. 938–943, 2003.

[12] "Louis vuitton," `http://www.louisvuitton.eu/front/\#/engE1/Collections/Women/Handbags/`.

[13] T. Stojic, I. Reljin, and B. Reljin, "Adaptation of multifractal analysis to segmentation of microcalcifications in digital mammograms," *Physica A: Statistical Mechanics and its Applications*, pp. 494–508, 2006.

[14] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun, "Salient object detection by composition," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1028–1035.

[15] T. Joutou and K. Yanai, "A food image recognition system with multiple kernel learning," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*, 2009, pp. 285–288.

[16] X. Xu, I. W. Tsang, and D. Xu, "Soft margin multiple kernel learning," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 24, no. 5, pp. 749–761, 2013.

[17] B. Yao, A. Khosla, and L. Fei-Fei, "Combining randomization and discrimination for fine-grained image categorization," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 1577–1584.

[18] A. Barla, F. Odone, and A. Verri, "Histogram intersection kernel for image classification," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, 2003, vol. 3, pp. III–513–16 vol.2.