On Branded Handbag Recognition

Yan Wang, *Student Member, IEEE*, Sheng Li, and Alex C. Kot, *Fellow, IEEE*

***Abstract*—Manufacturing branded handbags is a big business in the fashion world. Shoppers' feedback showing photos of their purchased handbags in social networks or blogs is important for branding purposes. In this paper, we deal with handbag recognition. It is a challenging problem due to the inter-class style similarity and the intra-class color variation. We focus on developing discriminative representations of handbag style and color. For handbag style representation, two supervised mid-level patch selection procedures are proposed to select discriminative patches, regarding individual classes and pairwise classes. We also propose a low-level complementary feature, extracted from texture-enhanced mid-level patches, to capture the fine details of the mid-level patches. For handbag color representation, we propose to extract dominant color features to handle the illumination changes. The performance of our proposed method is evaluated on a newly built branded handbag dataset. The results show that our method performs favorably in recognizing handbags, with around 10% improvement in accuracy when compared with the existing fine-grained or generic object recognition methods.**

***Index Terms*—Branded handbag recognition, discriminative, fine-grained object recognition.**

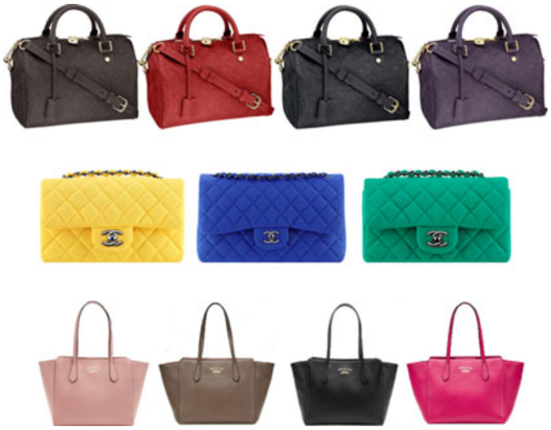


Fig. 1. Handbags with the same style but different models. From top to bottom, handbags of style name: Speedy Bandouliére from Louis Vuitton, Classic Flap Bag from Chanel, Swing Leather Tote from Gucci. We suggest that all the figures in this paper are best viewed in original color PDF file.

I. Introduction

THE explosive usage of personal devices with high resolution cameras have made visual object search more and more popular in our daily life, especially on the mobile end. A lot of research exists on visual object search, including user-centric mobile display [1], mobile video surveillance [2], multimedia retrieval [3]–[8], fashion search [9], fashion recommendation [10], [11], fashion parsing [12], and label or landmark recognition [13]. However, these works have not studied the problem of recognizing the handbag which has become an indispensable item of a person's wardrobe.

Sometimes consumers are attracted to a handbag posted either from fashion magazines or others' homepages (blogs, twitters or facebooks). They may only know the brand of this handbag based on its design or the description, but the specific model information is missing. Developing multimedia systems that automatically recognize the model name or number of a handbag is very important.

Image-based branded handbag recognition has practical potential in many applications, such as multimedia retrieval, fashion recommendation and fashion search. It is a new instance of fine-grained object recognition that is rarely studied before. Fine-grained object recognition concentrates on dealing with subtle differences between highly similar objects of the same category [14], [15]. A growing literature corpus has proposed various techniques for fine-grained object recognition. Some deal with domain-specific problems [15], [16], while others propose general frameworks which do not use strong class-specific knowledge i.e., information predicted on a specific object category [14], [17]–[20]. Template-based, patch-based and segmentation-based techniques [21]–[23] have been widely studied. However, these methods are rather limited when dealing with branded handbag recognition, due to handbag's inter-class style similarity and intra-class color variation. For simplicity, the "branded handbag" will be termed as "handbag" for short in the rest of the paper.

Each handbag is associated with a model number and a style name, which are provided by manufacturers. In general, handbags (with different models) sharing the same style name are different from each other only by their colors. Fig. 1 gives some examples for such handbags. A group of handbags only different in color are termed as a handbag Style-specific SubCategory (SSC) in our paper.

The challenges of handbag recognition lie in the following two parts:

1) Inter-class style similarity, where two major problems are needed to be solved for this challenge. Firstly, the styles of some handbag SSCs are very similar, only small decorations on local parts show discriminability, as shown in Fig. 2(a). Secondly, subtle texture difference between

Manuscript received July 21, 2015; revised November 19, 2015 and March 05, 2016; accepted May 18, 2016. Date of publication June 15, 2016; date of current version August 12, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Balakrishnan Prabhakaran.

Y. Wang and A. C. Kot are with the Rapid-Rich Object Search (ROSE) Laboratory, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 637553 (e-mail: wang0696@e.ntu.edu.sg; eackot@ntu.edu.sg).

S. Li is with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: shengli@shu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2581580

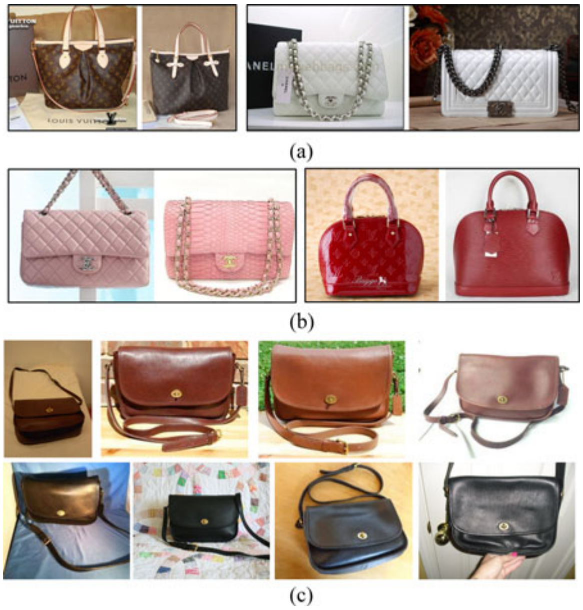


Fig. 2. Illustrations of the main challenges in handbag recognition. 1) Inter-class style similarity: in each black box of (a) and (b), we show two handbags of different SSCs, which only have subtle differences in style. 2) Intra-class color variation: in each row of (c), we show four handbag images belonging to the same model. The styles of two handbag models [the first and second rows in (c)] are the same, while their appearances differ in color.

handbag SSCs is non-trivial for the recognition. However, the texture is always not distinct due to the lighting condition or out of focus, such as the visually similar handbags but with different embossed texture patterns shown in Fig. 2(b).

2) Intra-class color variation: the illumination changes enlarge the intra-class color variance of handbag models within an SSC. Fig. 2(c) gives some examples for handbag models belonging to the same SSC.

In this paper, we aim at dealing with the aforementioned challenges. Based on the visual characteristics of handbags, we follow two directions to recognize the handbag model: 1) style-based SSC recognition and 2) color-based handbag model recognition. In the style-based SSC recognition, firstly, two supervised mid-level discriminative patch discovery strategies are proposed to separate visually similar SSCs. Secondly, we propose to extract low-level features (e.g., SIFT [24], HOG [25] or LBP [26]) from a texture enhanced handbag region and combine them with the features extracted directly from the original handbag region, so as to complement the details that are not captured by the latter. In this paper, we term the feature extracted from the original image, without any pre-processing of the image as original feature. The handbags within one SSC, are then recognized by using the dominant color feature elements, selected through a supervised manner.

As there is no benchmark available for handbag recognition, we build two branded handbag datasets for evaluation and we will make these datasets available in public. In terms of handbag recognition, our proposed method is shown to be able to achieve over 10% improvement in accuracy when compared with the existing fine grained or generic object recognition methods [17], [27], [28].

The rest of the paper is organized as follows. Section II briefly reviews the existing work on fine-grained object recognition and patch-based method for object recognition. Section III introduces our designed technique for handbag recognition. Section IV illustrates the dataset construction procedure and Section V presents the experimental results as well as some discussions, followed by the conclusion in the last section.

II. Related Work

A. Fine-Grained Object Recognition

Fine-grained object recognition is an emerging topic in recent years [15], [16], [22], [29]. There exist some famous benchmarks for different domains of fine-grained object recognition. CUB-200-2010 and CUB-200-2011 [30] are the two most competitive and challenging benchmarks for bird recognition. Oxford-IIIT-Pet dataset [15] is built for cat and dog recognition. The 102 Oxford flowers dataset [31] and the 578 flower dataset [32] are constructed for flower recognition. Different techniques were proposed to target on solving the fine-grained object recognition in different domains. For example, Branson *et al.* [16] exploited common visual characteristics of birds, coupled with interactive user input, to do the recognition. One of the most recent works about cats and dogs was presented by Parkhi *et al.* [15]. Pet face and fur are captured by a deformable part model and a bag-of-words model, respectively. For flower recognition, the works in [31] and [32] were proposed to allow part recognition and object segmentation to facilitate each other and further improve the classification performance. Some general methods were also proposed to solve general fine-grained object recognition problems. Local alignment-based methods [33], [34] roughly align the objects by the overall shape, and from which predicted parts or regions are derived. Deformable part descriptor-based methods [21] usually utilize object part annotations to localize the semantic parts. Detection and segmentation-based methods [22] aim to decrease the impact of background. Template-based methods [14], [23] seek out the right alignment of image regions that contain the same parts, where a large number of fixed-position templates are created to obtain image representations. Human-computer methods [19] require human interaction.

B. Patch-Based Method for Object Recognition

Patches have been sought as intermediate representations that can substitute lower level descriptors. Some previous research works deal with the problem of discovering discriminative patches for object recognition. Juneja *et al.* [35] proposed to learn discriminative part (patches) incrementally for scene classification. Endres *et al.* [36] learned a discriminative part collection to detect objects, which facilitates object recognition. Singh *et al.* [37] extracted mid-level discriminative patches, which can be used in a supervised regime to do the classification tasks.

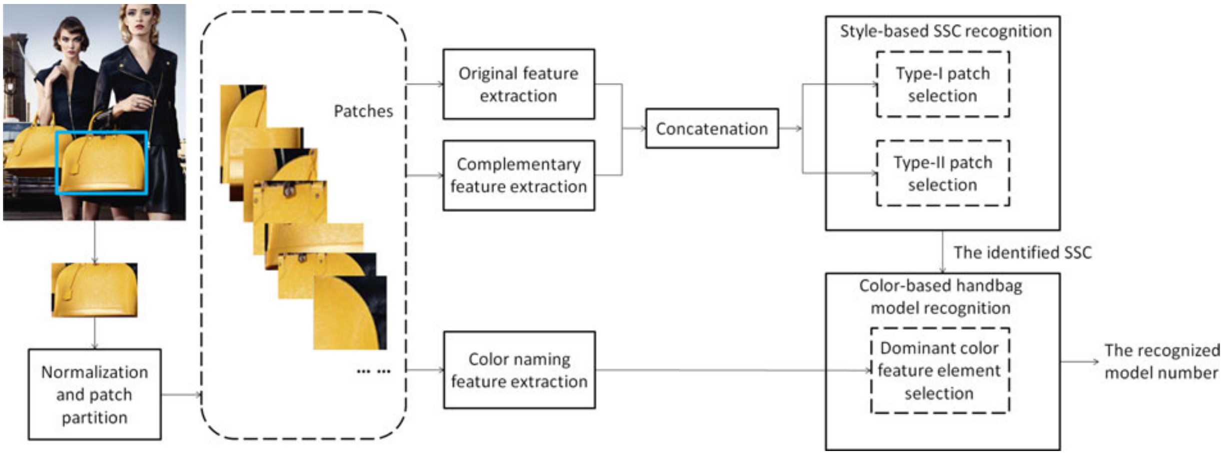

Patches
Original feature extraction
Complementary feature extraction
Concatenation
Style-based SSC recognition
Type-I patch selection
Type-II patch selection
The identified SSC
Color-based handbag model recognition
Dominant color feature element selection
The recognized model number
Color naming feature extraction
Normalization and patch partition
... ...

Fig. 3. Overview of the designed handbag recognition framework.

C. Instance Search

The objective of our task is to recognize a specific handbag, which is also related to the instance search problem, i.e., the problem of matching (recognizing) in images a particular object [38]–[40]. Over *et al.* [38] summarized and discussed the approaches and results of the TREC Video Retrieval Evaluation (TRECVID). To match objects, Li *et al.* [39] proposed an affine-invariant combination of each point in the model graph with its neighboring points. They built a fast linear matching framework which can handle complex transformations of an object. Then Li *et al.* [41] proposed to match each template feature point to all scene points based on their similarities, which improves the system's ability to deal with globally optimal transformation. Arandjelović and Zisserman [40] proposed to search for images containing one specific object from an image dataset by considering three aspects: descriptors, query expansion and feature augmentation.

The deformable part descriptor-based methods [21] require different visible parts or poses of the objects. Template-based methods [14], [23] demand a large number of training samples, while lacking the ability to generalize if the query image is different from the training data. Detection and segmentation-based methods [22] are not likely to perform well if the segmented region is not accurate. Patch-based methods are more suitable for our case but methods like that proposed by Yao *et al.* [17] sample patches densely from the image, which causes high computational cost. Instance search methods are effective and efficient to search for specific objects. However, the lack of supervision of handbag classes makes it difficult to differentiate visually similar handbags. In this paper, we propose discriminative representations for recognizing handbags. The discriminative patch serves as a supervised mid-level visual representation and the complementary feature serves as an unsupervised low-level visual representation for handbag style. We also select the most dominant color feature elements for representing handbag color.

III. Handbag Recognition

Fig. 3 illustrates the pipeline of the proposed handbag recognition method. Users first crop the handbag regions of interest (ROIs) from the images which they capture from fashion magazines or download from someone's blog. Then they upload the handbag ROIs to our system for further recognition. Each handbag ROI is rescaled to the uniform scale, and a set of K patches of multi-scales are extracted densely from it (see Section V for more details about the patch partition). A sequential method is applied to recognize the models of the input handbag ROIs. For recognizing the handbag style (i.e., to categorize the handbag ROI into an SSC), two supervised patch selection strategies are applied to capture the discriminative patches. From each patch a complementary feature is extracted and concatenated with the original feature to form a mid-level representation of the handbag. Then a dominant color feature element selection is applied for recognizing the handbag model of the input handbag ROI.

A. Style-Based SSC Recognition

The aim of this section is to identify which SSC a query handbag belongs to. The key problem is to find and make use of the discriminative handbag patches for recognition. We propose two patch selection mechanisms, which are to select the discriminative patches w.r.t. one individual SSC and pairwise SSCs, respectively.

1) Patch Selection for Individual SSC: The texture difference among handbag SSCs can be very subtle and only some patches matter. It is necessary to find the discriminative patches that can capture the intrinsic characteristics of each SSC. To this end, we propose a random forest-based strategy [42] to measure how discriminative a patch is for identifying handbag SSCs.

Each random forest tree consists of several branch nodes and leaf nodes, as illustrated in Fig. 4(a). In each branch node, the input training samples (represented by features) Q are split into left and right subsets Q_l and Q_r respectively, as shown in Fig. 4(b). The binary split is determined by a pair of well chosen parameters $\phi^* = (\theta^*, \tau^*)$, where θ^* is the index of the most discriminative (i.e., θ^*th) feature element and τ^* is the corresponding threshold. Usually the quality of the split can be measured by using the reduction of Gini impurity [43] before and after partition. Here, we follow the same procedure in the standard random forest pipeline [44], [45] for selecting the most

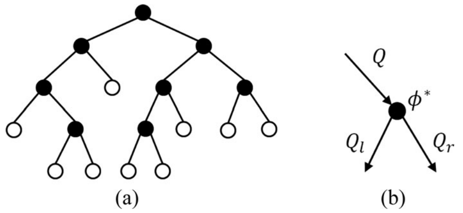


Fig. 4. Structure of a random decision tree. A forest is an ensemble of trees. (a) Each tree consists of branch nodes (filled circle) and leaf nodes (empty circle). (b) For each branch node, a set of input training samples Q are needed to be binary splitted into left and right subsets Q_l and Q_r.

discriminative feature elements and storing their indexes. We propose to measure the discriminability of a patch based on its feature elements that are selected among all the nodes.

Let's denote the feature of the kth ($k = 1, 2, \ldots, K$) partitioned patch for a handbag ROI r as $\mathbf{f}(r(k))$, a W dimensional vector. By concatenating the features of all the K patches, we obtain a $(K \times W)$ dimensional vector for each handbag ROI. We then propose to measure the discriminability of the kth patch by

$$d(k) = \sum_{t=1}^{T} \Phi(t, k) \tag{1}$$

where T is the number of the branch nodes for all trees and

$$\Phi(t, k) = \begin{cases} 1, & \text{if } \theta^*(t) \in [(k-1)W + 1, kW] \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\theta^*(t)$ refers to the index of the feature element chosen in the tth node. The Y most discriminative patches obtained according to (1) (i.e., the patches with the Y largest $d(k)$) are selected for future recognition.

2) Patch Selection for Pairwise SSCs: The aforementioned patch selection considers the overall discrimination of the patches for an individual SSC. However, the patches that best differentiate the two similar SSCs may not be treated as discriminative patches to classify other SSCs. To enable our style-based SSC recognition to capture the differences between SSCs with very similar styles, we propose to find the most discriminative patch for a pair of SSCs. The location of the most discriminative patches for different pairs of SSCs may vary, as shown in Fig. 5.

Assume that there are a pair of SSCs used for training: $A = \{a_i\}_{i=1}^{n}$ and $B = \{b_j\}_{j=1}^{m}$, where a_i and b_j denote the normalized handbag ROIs from SSC A and SSC B, respectively. To simplify the symbol, we drop the subscript i, j and define $d_{a,b}(k)$ as the Chi-square distance between the kth ($k = 1, \ldots, K$) patches in normalized handbag ROIs a and b

$$d_{a,b}(k) = \chi^2 \left(\mathbf{f}(a(k)), \mathbf{f}(b(k))\right). \tag{3}$$

We then define the discriminability of the corresponding kth patch for a pair of SSCs (A, B) as

$$D_{A,B}(k) = \frac{d_{A,B}(k)}{\left(d_{A,A}(k) + d_{B,B}(k)\right)/2} \tag{4}$$

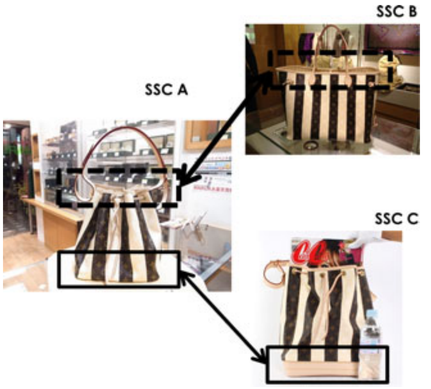


Fig. 5. Examples for the most discriminative part of pairwise SSCs. The dotted rectangles (upper region) are the most distinctive part for SSC A and SSC B; the rectangles with solid lines (lower region) are the most distinctive part for SSC A and SSC C.

where

$$d_{A,B}(k) = \frac{1}{n \times m} \sum_{\substack{i=1,\ldots,n,\\ j=1,\ldots,m}} \left(d_{a_i,b_j}(k)\right) \tag{5}$$

$$d_{A,A}(k) = \frac{1}{n \times (n-1)} \sum_{\substack{i,i'=1,\ldots,n,\\ i \neq i'}} \left(d_{a_i,a_{i'}}(k)\right) \tag{6}$$

$$d_{B,B}(k) = \frac{1}{m \times (m-1)} \sum_{\substack{j,j'=1,\ldots,m,\\ m \neq m'}} \left(d_{b_j,b_{j'}}(k)\right). \tag{7}$$

For the kth pair, (5) measures an inter-class distance between SSC A and SSC B; (6) and (7) measure the intra-class distances of SSC A and SSC B respectively. Suppose there are in total N SSCs for a certain brand, for ith ($i = 1, \ldots, C_N^2$) pair of SSCs, we choose the $Z(i)$th patch among the K pairs of corresponding patches as the most discriminative patch for future recognition

$$Z(i) = \arg\max_{k} \left\{D_{A,B}(k)\right\}. \tag{8}$$

Now that we have selected two types of discriminative patches to facilitate the SSC recognition, we term these two types of discriminative patches as the following.

1) Type-I patch: one of the Y discriminative patches selected for an individual SSC.
2) Type-II patch: the most discriminative ($Z(i)$th) patch for ith pair of SSCs.

Among N SSCs, first of all, we concatenate the features extracted from the Type-I patches and form a feature vector $\mathbf{g}$. To distinguish the ith pair of SSCs, we train a binary classifier $R(i)$ based on the features extracted from the $Z(i)$th patch of the pair of SSCs, where $Z(i) = 1, 2, \ldots, K$ is the Type-II patch of this pair of SSCs. Since the number of $R(i)$s is somewhat large (in total C_N^2 binary classifiers), we adopt the fast and effective logistic regression formulation[1] for the training. After each $R(i)$ has been trained, we extract the feature of the $Z(i)$th patch of

[1] [Online]. Available: http://ufldl.stanford.edu/wiki/index.php/Softmax_Regression/

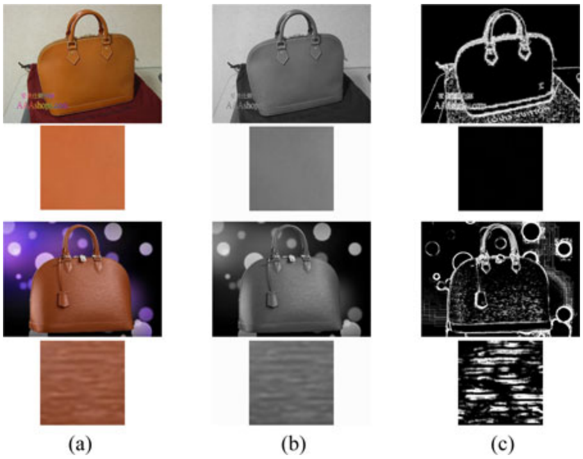


Fig. 6. Two handbags with similar styles. (a) Original handbag images. (b) Gray images. (c) α-images.

each handbag ROI and feed it to $R(i)$ to obtain a score. We then concatenate all the scores (in total C_N^2 scores) and form a feature vector $\mathbf{g}'$. The final feature vector $\mathbf{h}$ is the concatenation of $\mathbf{g}$ and $\mathbf{g}'$, which is adopted to train an N-class SVM classifier C for SSC recognition.

B. Complementary Feature Extraction

We can use any of the existing features developed in [24]–[26], [46] to represent handbag patches. However, we find that the subtle difference among different handbags may not be shown prominently in their original images due to the lighting conditions or out of focus. Fig. 6 shows two different handbags. It is difficult to tell the difference between these two handbags from the original images or gray images (see Fig. 6(a) and (b), first and third rows). If we look into the zoomed in patches (see Fig. 6(a) and (b), second and fourth rows), the details of the textures are shown but still not prominent. Therefore, applying the original feature may not capture sufficient details of the handbag textures.

In this section, we extract the feature from the texture enhanced image, which is termed as the complementary feature [47], to complement the details of the texture that are not captured by the original feature. The enhancement is performed by using Hölder exponent from the multifractal theory [48]. As indicated in [48], Hölder exponent has the ability to enhance the image local structure. We denote S_ε as a squared local image block, where ε refers to the number of pixels along one side. According to [48], the value of Hölder exponent α is estimated as a slope of linear regression line in a discrete space. Slightly different from [48], to efficiently capture the regularity of the local structure for a pixel located at (i, j) in the image, we consider its nearest neighborhood up to $\varepsilon = 3$ (i.e., $\varepsilon = 3$ and 1). Thus the Hölder exponent for this pixel is computed as

$$\alpha(i,j) = \frac{\ln \hbar(S_3(i,j)) - \ln \hbar(S_1(i,j))}{\ln(3) - \ln(1)} \tag{9}$$

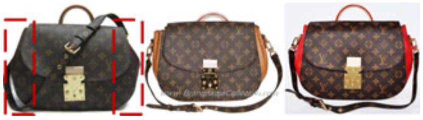


Fig. 7. Examples of different handbag models in one SSC. The color from certain patches is more discriminative (in dotted boxes).

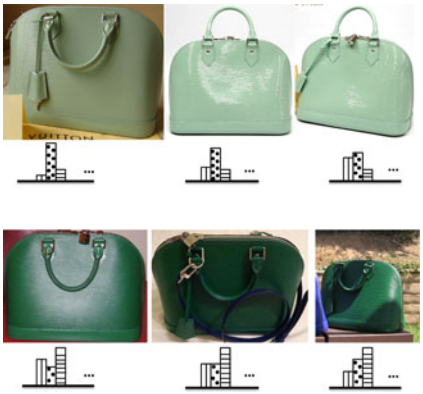


Fig. 8. Handbags with the corresponding color histograms. The first row and third row are two handbags with three different images per handbag model. The second row and fourth row are corresponding extracted color histograms.

where $S_\varepsilon(i,j)$ refers to the $\varepsilon \times \varepsilon$ block with the center located at (i, j). In our implementation, $\hbar(S_\varepsilon(i,j))$ is the maximum pixel value within $S_\varepsilon(i,j)$.

$\alpha(i,j)$ can be normalized to $\alpha^*(i,j) \in [0, 255]$ to form a grayscale image α^*, which is termed as α-image [see Fig. 6(c)]. Observe that the details of handbag textures are enhanced in the α-images, which provide more discriminative information to recognize different handbags. After texture enhancement, the complementary feature can be extracted from the α-image of the handbag using existing feature extractors such as SIFT, HOG, etc., which can further improve the discriminative power of mid-level handbag patches. We are the first to incorporate α-image with those feature descriptors to improve the classification performance.

C. Color-Based Handbag Model Recognition

In this section, we proceed to further recognize the handbag model within each SSC based on the color information only. As discussed before, handbags within one SSC differ from each other just by their colors. However, to differentiate handbag models in one SSC, extracting an F dimensional color histogram from ROIs is not discriminative enough for two reasons:

1) Sometimes only the color of some corresponding patches are different, as shown in Fig. 7;
2) Sometimes the color histograms for images belonging to the same handbag model vary a lot because of the diverse illumination environments, such as handbags with glossy material, as shown in Fig. 8. However, a few color elements may remain consistent with each other (see the third color element shown with horizontal textures in Fig. 8).

Therefore, we extract an F dimensional color histogram for each of the K patches from a handbag ROI. Among all these features, discriminative color elements are extracted and termed as the dominant color feature elements.

To describe the color information of a handbag region, we concatenate F dimensional color histogram extracted from each handbag patch, and form an $(F \times K)$ dimensional color feature. We obtain the dominant color feature elements for a certain SSC by following a similar procedure to the Type-II patch selection described in the previous section. Recalling that SSC $A = \{a_i\}_{i=1}^{n}$ denotes an SSC, it can be rewritten as $A = \{H_l\}_{l=1}^{L}$, where L denotes the number of handbag models in A and H_l is the set of the normalized handbag regions belonging to lth handbag model. H_l^* denotes the set of handbag regions which do not belong to the lth handbag model. Let $q(a)$ be the mapping function between a normalized handbag region a and its model index, i.e., $q(a) \in \{1, \ldots, L\}$, then we can obtain $H_l = \{a_i | q(a_i) = l\}_{i=1}^{n}$. Let $|\cdot|$ be the number of samples in a set, $h_l^* = \{h_t | t = 1, \ldots, L, t \neq l\}$, and $\bar{d}_{a_i,a_j}(k)$ denotes the Euclidean distance between kth ($k = 1, \ldots, F \times K$) color dimension of normalized handbag regions a_i and a_j ($a_i, a_j \in A$). The discriminability of the kth color dimension for SSC A is computed by

$$D_A(k) = \frac{1}{L} \sum_{l=1}^{L} \frac{\bar{d}_{H_l,H_l^*}(k)}{\left(\bar{d}_{H_l,H_l}(k) + \bar{d}_{H_l^*,H_l^*}(k)\right)/2} \tag{10}$$

where

$$\bar{d}_{H_l,H_l^*}(k) = \frac{1}{|H_l| \times |H_l^*|} \sum_{\substack{q(a_i)=l,i=1,\ldots,n,\\ q(a_j)\neq l,j=1,\ldots,n}} \left(\bar{d}_{a_i,a_j}(k)\right) \tag{11}$$

$$\bar{d}_{H_l,H_l}(k) = \frac{1}{|H_l|^2 - |H_l|} \sum_{\substack{q(a_i)=l,q(a_{i'})=l,\\ i,i'=1,\ldots,n,i\neq i'}} \left(\bar{d}_{a_i,a'_i}(k)\right) \tag{12}$$

$$\bar{d}_{H_l^*,H_l^*}(k) = \frac{1}{|H_l^*|^2 - |H_l^*|} \sum_{\substack{q(a_i)\neq l,q(a_{i'})\neq l,\\ i,i'=1,\ldots,n,i\neq i'}} \left(\bar{d}_{a_i,a_{i'}}(k)\right). \tag{13}$$

For the kth color dimension, (11) measures the distance between handbags which belong to lth model and those belong to other models (i.e., inter-class distance). (12) and (13) are the normalized factors, and they measure the distance of handbags which belong to the lth model and other models, respectively (i.e., intra-class distance). We rank each dimension according to its discriminability computed by (10) in a descending order. First P dimensional color feature elements with the highest discriminability are chosen as the dominant color feature elements. For handbag model(s) of an SSC, if $L > 1$, we further train an L-class classifier based on the dominant color feature elements. For efficiency, we adopt the fast and effective multi-class softmax regression (see footnote 1) for the handbag model recognition.

IV. Dataset Construction

As no existing handbag dataset is available for handbag recognition, we construct two datasets consisting of over 5000 images from 220 handbag models of one brand[2] and over 4000 images from 181 handbag models of another brand.[3] We name these datasets as BrandBag-I and BrandBag-II, respectively, and name the combination of two brands as BrandBag. Each image in our dataset is annotated with its handbag model label, style name label, and a bounding box indicating its ROI.

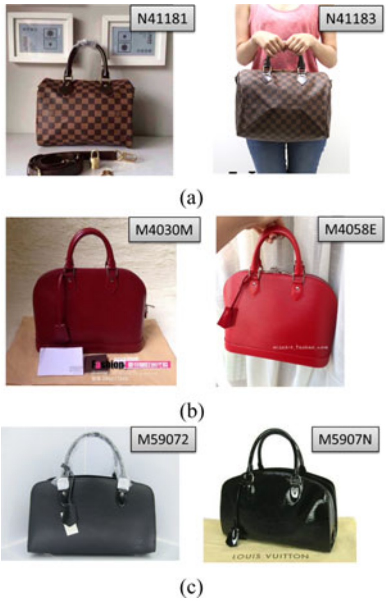


Fig. 9. Examples of visually indistinguishable handbags. The handbags have the same appearance but have (a) different sizes, (b) indistinguishable colors, and (c) different materials. The number associated with each handbag refers to the corresponding model.

A. Handbag Image Collection

We ask several subjects (with different nationalities, and age ranges from 17–36) to construct the datasets. First, the list of target handbag models is obtained from the websites (see footnote 2 and 3), each of which is associated with a style name. Then, images are collected using Google and Flickr image search by typing the keywords like the brand with the model number. Some handbag models have the same appearance but different sizes, indistinguishable colors, or different materials, as shown in Fig. 9. Such handbags can not be distinguished by images. Therefore, they are merged as one model. To ensure enough samples for training and testing, we exclude the handbags with only a few images available (fewer than 10). Eventually, the datasets contains 5545 images of 220 BrandBag-I handbag models and 4318 images of 181 BrandBag-II handbag models.

B. SSC Grouping

For handbags inside one brand, those only different with each other in color are grouped as one SSC. The grouping procedure consists of three steps.

[2] "Louis Vuitton," [Online]. Available: http://www.louisvuitton.eu/front/#/eng E1/Collections/Women/Handbags

[3] "Coach," [Online]. Available: http://www.coach.com/shop/women-handbags?viewAll=true

TABLE I
DISTRIBUTION OF NUMBER (#) OF HANDBAGS PER SSC

# of handbags	# of SSCs	
per SSC	BrandBag-I	BrandBag-II
1	74	23
2	33	9
3	10	9
4	4	9
5	2	5
6	0	0
7	0	2
8	0	1
9	0	1
10	0	1
11	0	1
12	1	0
13	1	0

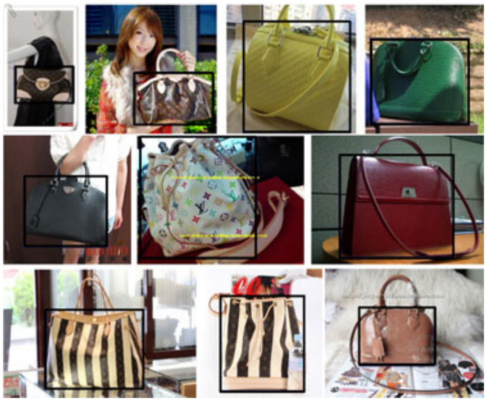


Fig. 10. Examples of handbag images with the associated bounding boxes (marked with black rectangles) in our datasets.

1) For each handbag dataset, handbags with the same style name are grouped first.
2) For most cases, a style name group is regarded as an SSC if handbags with this style name differ from each other only in color.
3) For only a few cases, subjects divide a style name group into more SSCs if besides color, handbags with the same style name have some difference.

Eventually, our BrandBag-I dataset consists of 125 SSCs and the number of handbags per SSC is in the range of $[1, 13]$. The BrandBag-II dataset consists of 61 SSCs. The number of handbags per SSC is in the range of $[1, 11]$. More specifically, the distribution of number of handbags per SSC in our datasets is illustrated in Table I.

C. Bounding Box Annotation

Each handbag image in our dataset is annotated with a bounding box covering the handbag surface, as shown in Fig. 10. We ask at least six subjects to annotate one image. An interface is built to facilitate the cropping procedure. The cropping is done by dragging a rectangle on the screen from top left to bottom right of the handbag. After removing the bad rectangles which is far from fitting around the handbag region, the final bounding box is constructed by taking the average locations of each side of the cropped rectangles.

The difficulties for collecting the datasets are listed below. For many handbag models, it is difficult to collect the images from online sources, thus we expand the query keywords and search the images through multiple search engines. The retrieved handbag images from the internet sometimes do not match the specified style name or model number, thus we use the images downloaded from (see footnote 2 and 3) as the references when refining the dataset.

Handbag images in our dataset are mainly in frontal view or with small rotations (i.e., the angle of the rotation is smaller than 30°), which is the best view to visualize handbags. Handbags appearing in these images are without heavy occlusion (i.e., the occluded region is no larger than 15% of the handbag size). The sizes of the images range from 93×99 to 3648×3968, while the sizes of the handbags in images are from 50×96 to 2025×3707. To the best of our knowledge, these are the first handbag datasets constructed for branded handbag recognition.

V. EXPERIMENTS AND DISCUSSIONS

A. Generality Evaluation for Complementary Feature and Patch Selection

1) Evaluation of the Complementary Feature: We extend our work in [47] to investigate the generality of the complementary feature based on the Locality-constrained Linear Coding (LLC) framework [27] on some famous benchmarks. They include five widely used image classification datasets: Caltech-101 [49], Caltech-256 [50], MIT 67 indoor [51], Scene 15 [52], UIUC sports [56] and some fine-grained object datasets: Oxford flower [31], Stanford dog [54] and UCSD bird [58]. We also test the complementary feature based on an existing fine-grained object recognition method for bird recognition [17]. Gray scale information is used for all these datasets.

We test all datasets with the most standard settings. On Caltech-101, we randomly selected 15 and 30 training images per class and no more than 50 testing images per class. On Caltech-256, 15, 30 and 45 images per class are randomly selected for training and the rest for testing. On MIT 67 indoor, we follow the original splits [51], which use around 80 training images and 20 testing images per class, respectively. On Scene 15, 100 images per class are randomly selected for training and the rest are for testing. On UIUC sports, we randomly select 70 training images and 60 testing images per class, respectively. On Oxford flower, whole images are applied for training and testing, while for Stanford dog, only foreground images are adopted. We follow the data preparation of UCSD bird in the work of Yao *et al.* [17]. We trained codebooks with 4096 bases for Caltech-256 as in [27] and 2048 for other datasets. We then adopt Spatial Pyramid Matching (SPM) [52] pooling to get the final representation, and followed by training an SVM classifier. We adopt the default settings in the source code [27], and extract SIFT [24] feature as the original feature. For simplicity, the corresponding complementary feature is termed as Hölder-SIFT. As shown in Table II, incorporating the complementary feature on general

TABLE II
MEAN CLASSIFICATION ACCURACY (%) OF WITH/WITHOUT THE COMPLEMENTARY FEATURE FOR IMAGE CLASSIFICATION

Datasets	Features	
(Number of training images/class)	SIFT	SIFT & Hölder-SIFT
Caltech-101 (15)	63.82	68.43
Caltech-101 (30)	71.93	76.27
Caltech-256 (15)	30.29	33.78
Caltech-256 (30)	36.77	40.63
Caltech-256 (45)	40.13	44.09
MIT 67 indoor (80)	41.51	46.57
Scene 15 (100)	81.65	84.66
UIUC sport (70)	81.50	87.71
Oxford flower (10)	42.38	48.88
Stanford dog (105)	20.76	24.08
UCSD bird (15)	12.71	15.73

TABLE III
MEAN CLASSIFICATION ACCURACY (%) OF WITH/WITHOUT COMPLEMENTARY FEATURE ON THE EXISTING FINE-GRAINED RECOGNITION METHOD [17] FOR BIRD RECOGNITION

	Feature	
Number of decision trees	SIFT	SIFT & Hölder-SIFT
50	11.97	13.74
100	13.49	15.97
150	14.55	16.83
200	15.22	17.64
250	15.43	18.00
300	15.68	18.32

object recognition helps in increasing the performance by over 3% in mean classification accuracy.

For bird recognition proposed by Yao *et al.* [17], we adopt the default settings in the source code, and change the number of trees and size of images given in the paper. The results in Table III show the improvement of incorporating complementary feature.

Besides, we test the complementary feature based on the LLC framework [27] on our handbag datasets. Five images per handbag model are randomly chosen for training and the rest are for testing. The training and testing sets are fixed for the experiments. Different types of features extracted from handbag ROIs are compared during the recognition, i.e., the original feature, the complementary feature, and the feature concatenated by the original and the corresponding complementary feature. We apply several popular features including SIFT [24], HOG [25], LBP [26] and Texton [46] as the original features. The corresponding complementary features are Hölder-SIFT, Hölder-HOG, Hölder-LBP and Hölder-Texton, respectively. Table IV shows the comparison results of accuracy. The concatenation of the original feature and the complementary feature is denoted as SIFT & Hölder-SIFT, HOG & Hölder-HOG, etc. We observe that most of the complementary features perform worse than the original features because they enhance high-frequent information which is sensitive to noise. However, concatenating the original feature and the corresponding complementary feature can get a better performance compared with only using the original feature. To demonstrate the effectiveness of extracting the feature from our texture enhanced image, we compare it with sketch token [56], which is another local edge-based feature. Likewise, we denote the corresponding features extracted from sketch tokens by ST-SIFT, ST-HOG, ST-LBP and ST-Texton. The 5th to the 8th columns and the 3rd to the 6th rows of Table IV show the performance of concatenating the edge-based feature and the original feature, which does not help in boosting the accuracy.

Since among different original features (the 1st to the 4th columns and the 3rd to the 6th rows in Table IV), use SIFT alone can achieve the highest accuracy in handbag recognition for BrandBag-I and BrandBag datasets. Next, SIFT will be adopted to test our methods for handbag recognition, regarding the patch partition and the patch selection. Then we will use SIFT & Hölder-SIFT to evaluate the overall framework.

2) Evaluation of the Patch Selection: To investigate the generality of the proposed patch selection procedure, we test it on UCSD bird dataset [55]. As bird does not have SSC, we apply the patch selection technique (feature $\mathbf{g}$, $\mathbf{g}'$ or the concatenation of $\mathbf{g}$ and $\mathbf{g}'$ illustrated in Section III-A.2 followed by SVM) directly on different bird categories. Patches are extracted randomly from the image. LLC features [27] coded from the densely sampled SIFT descriptors, followed by max-pooling are extracted to represent each patch. The codebook size is set as 256, which is consistent with the UCSD bird settings in [17]. The mean classification accuracies by using the proposed Type-I, Type-II and Type-I + Type-II for bird recognition are summarized in Table V (see the second column). Compared with adopting the LLC framework (with codebook size as 256) and Yao's method (see Table III), the proposed patch selection performs better. We further incorporate the complementary feature, and the mean classification accuracies for different methods are shown in Table V (see the third column).

B. Evaluation of the Proposed Framework for Handbag Recognition

In this section, we evaluate the performances of the handbag recognition. In the following discussions, we denote SIFT + LLC to represent the LLC features [27] coded from the densely sampled SIFT descriptors followed by max-pooling [59]. The size of the codebook equals to 1000.

Two baselines considered for comparison are as follows:

1) Baseline-I: C-SIFT [60] + LLC for handbag model recognition.
2) Baseline-II: we sequentially recognize the handbag from SSC recognition to model recognition. SIFT + LLC is used for SSC recognition, and the color naming [61] is adopted for model recognition because color naming has been found to be a successful color feature for image classification [62].

Each image is divided into 1 × 1, 2 × 2, and 4 × 4 spatial pooling regions [52] for C-SIFT + LLC in Baseline-I and SIFT + LLC in Baseline-II. We concatenate and normalize the pooled features from each region, and regard them as the final image

TABLE IV
ACCURACY (%) OF LLC FRAMEWORK ON HANDBAG RECOGNITION USING DIFFERENT TYPES OF FEATURES

	Datasets				Datasets		
Features	BrandBag-I	BrandBag-II	BrandBag	Features	BrandBag-I	BrandBag-II	BrandBag
SIFT	70.02	34.66	54.81	SIFT & ST-SIFT	70.09	34.57	54.95
HOG	63.99	29.27	48.69	HOG & ST-HOG	63.36	28.86	48.08
LBP	48.81	24.88	37.65	LBP & ST-LBP	46.67	24.89	37.46
Texton	68.38	35.01	53.39	Texton & ST-Texton	68.29	35.42	53.29
Hölder-SIFT	67.43	31.41	52.11	SIFT & Hölder-SIFT	72.77	37.06	57.44
Hölder-HOG	60.68	27.16	46.33	HOG & Hölder-HOG	67.45	31.06	51.55
Hölder-LBP	48.09	19.22	35.01	LBP & Hölder-LBP	52.23	26.23	40.63
Hölder-Texton	68.81	26.96	50.25	Texton & Hölder-Texton	73.78	36.98	57.56

TABLE V
COMPARISONS OF THE MEAN CLASSIFICATION ACCURACY (%) FOR BIRD RECOGNITION

	Feature	
Methods	SIFT	SIFT & Hölder-SIFT
LLC framework	11.64	16.27
Type-I	14.88	17.38
Type-II	15.07	17.02
Type-I + Type-II	16.97	18.82

feature representation. Then we train SVM classifiers based on the feature representation. For model recognition in Baseline-II, the multi-class softmax regression (see footnote 1) is adopted. The classification accuracies (when compare the ground truth against the predicted class across all test images) of the two baselines on our dataset are shown in Table VI. It can be seen that the sequential recognition process can significantly increase the performance for handbag recognition.

Next, we are to evaluate the performance of proposed discriminative representations for handbags. Each handbag ROI is rescaled into 255×255 pixels. We observe that handbag designers tend to decorate handbags with patterns or textures in the four corners, middle, upper or lower part, and the two sides. This motivates us to partition each ROI into scales of 85×85, 85×170, 170×85, 85×255 and 255×85 with a step size of 85, which obtains $K = 27$ patches in our experiment. To categorize the handbag ROI into an SSC, SIFT + LLC is adopted to represent patches unless otherwise specified. To further recognize the model of the ROI, the color naming method [61] is adopted to represent patches.

1) Performance of Type-I Patch Selection: We use the selected Type-I patches instead of the standard SPM patches in Baseline-II pipeline and vary the number of selected patches. We select $Y = 15$ most discriminative patches by grid search for the datasets. The result is shown as Type-I + color naming in Table VI. If we use all the 27 patches for handbag recognition, the accuracies are 86.17%, 66.58% and 77.91% for BrandBag-I, BrandBag-II and BrandBag, respectively. Moreover, we observe that the performances are not sensitive to $Y \in [13, 18]$, where accuracies are within the range of $[88.09\%, 88.54\%]$, $[67.65\%, 68.06\%]$ and $[79.35\%, 79.60\%]$ for the three datasets, respectively. Our Type-I patch selection is more efficient compared with SPM, as it can select less patches while achieving better results.

We also visualize the four most discriminative patches as well as the four least discriminative patches according to (1) in Fig. 11. Here BrandBag-I dataset is taken as an example. The ranking shows that the discriminative patches locate on the upper and middle part of the handbags in the dataset.

2) Performance of Type-II Patch Selection: Fig. 12 shows the ranking of discriminability of patches for pairwise SSCs in a descending order. The top-ranked patch indicates that the diversity of bottom corner is important for distinguishing these two SSCs.

As discussed in Section III-A.2, among N SSCs, C_N^2 binary classifiers will be built in this module to construct the feature $\mathbf{g}'$. There are 125 SSCs in BrandBag-I, 61 SSCs in BrandBag-II and 186 SSCs in BrandBag, yielding

$$\binom{125}{2} = 7750, \binom{61}{2} = 1830$$

and

$$\binom{186}{2} = 17205$$

dimensions for $\mathbf{g}'$, respectively. However, it is not necessary to use all the elements in $\mathbf{g}'$. Table VII shows the performance of handbag recognition with $\mathbf{g}'$ by using randomly selected number of pairwise classifiers (i.e., $R(i)$) on top of the Type-I patch selection with $Y = 15$. We denote the number of selected pairwise classifiers as E and we are able to gain a good performance when E is small. For example, for BrandBag-I, this is around 2% improvement compared with only using the Type-I patches with $Y = 15$ (accuracy $= 88.54\%$). We also report the performance of adopting only Type-II patches for the style representation, shown as Type-II + color naming in Table VI.

3) Comparisons of Discriminative Patch Discovery: We replace the proposed patch selection with the leading and publicly available method [17] for fine-grained object recognition (see Yao *et al.* [17] + color naming in Table VI), discriminative patch discovery method [37] (see Singh *et al.* [37] + color naming in Table VI) and Feng *et al.* [57] (see Feng *et al.* [57] + color naming in Table VI). We adopt the default settings in their source

TABLE VI
COMPARISONS OF THE CLASSIFICATION ACCURACIES (%)

	Datasets		
Methods	BrandBag-I	BrandBag-II	BrandBag
Baseline-I	67.21	39.00	54.08
Baseline-II	81.33	60.62	72.83
Type-I + color naming	88.54	67.86	79.59
Type-II + color naming	87.95	67.57	79.08
Type-I + Type-II + color naming	90.77	70.91	82.19
Yao *et al.* [17] + color naming	80.11	66.74	74.01
Singh *et al.* [37] + color naming	71.04	39.68	64.03
Feng *et al.* [57]	49.44	41.80	45.19
Type-I + Type-II + complementary feature + color naming	92.34	71.96	83.46
Type-I + Type-II + complementary feature + color selection (Proposed)	92.77	72.25	84.01
Yao *et al.* [17]	76.28	39.56	61.47
Vedaldi & Zisserman [58]	40.99	11.43	19.75
Convolutional Neural Network (CNN) [28]	80.18	60.69	71.27
CNN_6 feature	86.04	61.38	75.40
CNN_7 feature	85.79	62.14	74.76
Type-I + Type-II (CNN_6) + color selection (Proposed)	89.98	69.91	83.79

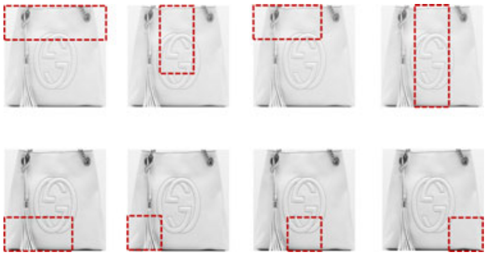


Fig. 11. Patches arranged by their discriminability in a descending order for BrandBag-I. We visualize the most four (first row) and the least four (second row) discriminative patches (patches are indicated as dotted boxes).

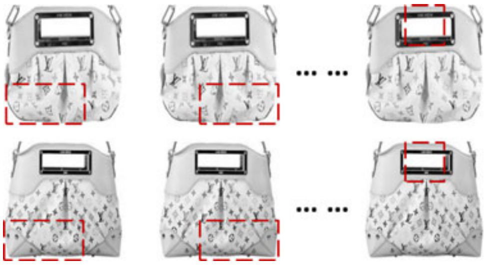


Fig. 12. An example of learned discriminative patches for two SSCs (shown in two rows respectively). We rank the discriminability of patches in a descending order, and visualize the two most discriminative patches (in dotted boxes) as well as the least discriminative patch.

codes. For Singh's work, we discover discriminative patches on a per-category basis, and aggregate top discovered patches of each SSC into an object bank representation [62]. For Feng's work, we regard the top ten salient windows extracted from each handbag ROI as discriminative patches. Our patch selection for the task achieves a large improvement in accuracy. Previous methods [17], [37], [57] attempt to find patches of the image that are discriminative without explicit part locations. The use of randomization leads to some information loss for specific patches, while our method introduces domain-specific prior into the discriminative patch localization.

4) *Performance of Complementary Feature Extraction:* We concatenate SIFT + LLC with Hölder-SIFT + LLC to represent handbag patches in Type-I + Type-II + color naming. The accuracy is reported as Type-I + Type-II + complementary feature + color naming in Table VI.

5) *Performance of Dominant Color Feature Element Selection:* As mentioned in Section III-C, we choose the P most discriminative color elements (out of $11 \times 27 = 297$ dimensional feature extracted from all patches) for the color-based handbag model recognition. We are able to achieve a high accuracy at $P = 230$ for handbag recognition (on top of the Type-I and Type-II patch selection, based on grid search). In our dataset, the number of handbags in each SSC is relatively small. This may be one of the reasons why the improvement is little compared with using all the color feature elements.

TABLE VII
VARIATION WITH SELECTED NUMBER OF PAIRWISE CLASSIFIERS (E) FOR HANDBAG RECOGNITION IN ACCURACY (%)

E	BrandBag-I	E	BrandBag-II	E	BrandBag
2000	90.77	400	69.06	4000	82.19
4000	90.74	800	70.52	8000	82.21
6000	90.81	1200	70.73	12000	82.24
7750	90.81	1830	70.91	17205	82.17

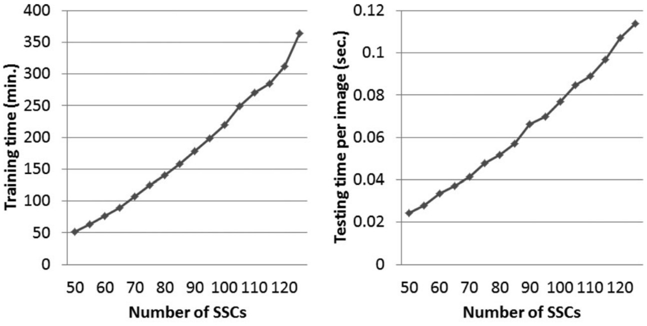
Finally, we compare our proposed framework with some famous methods. We directly adopt Yao's method for recognizing handbag models, denoted as Yao *et al.* [17] in Table VI. We also compare with a baseline for instance search in [58], shown as Vedaldi & Zisserman [58]. As Convolutional Neural Network (CNN) has been shown to be effective for many image recognition tasks, we adopt the ImageNet pre-trained model [28] and fine-tune it on our handbag data. We start the training with a fixed learning rate and decrease it by a factor of 10 after the training

TABLE VIII
CLASSIFICATION ACCURACIES (%) BY USING THE AUTOMATIC SSC PROCESS WITHOUT (W/O) AND WITH (W/) COLOR INFORMATION

	Datasets		
Methods	BrandBag-I	BrandBag-II	BrandBag
w/o color information	90.65	46.50	65.40
w/ color information	90.86	65.66	80.15

error stops reducing, and the result is reported as CNN [28] in Table VI. CNN is also extensively used for feature extraction, we extract the CNN feature for each patch via the fine-tuned CNN network, and use the outputs from the 6th layer (CNN_6) or the 7th layers (CNN_7) [64]. We use the concatenated CNN features of all patches and followed by an SVM classifier for handbag model recognition. The results are shown as CNN_6 feature and CNN_7 feature in Table VI. We also replace the concatenated feature of SIFT + LLC and Hölder-SIFT + LLC with CNN_6 feature for style representation in our framework. Table VI shows that adopting the CNN feature can not improve the recognition performance. Compared with other frameworks, our proposed framework achieves around 10% improvement in classification accuracy. The strength of our method lies in its discriminative representation which aims at dealing with difficulties in handbag recognition. Compared with CNN, we explicitly design mid-level discriminative patches with strong location information. We also design a low-level complementary feature to capture the subtle difference among SSCs with a small number of training data. Our style-to-color discriminative representation strategy takes the advantages of visual characteristics of handbags. It obeys the rule of handbag design. Our framework works well for most handbag models. However, the handbag images of some models suffer from severe illumination and non-rigid deformation, which results in a decrease in classification accuracy. For color-based handbag model recognition, although we propose to select the dominant color feature elements to represent the handbag, our algorithm relies on the quality of the original color feature.

We report the results of an automatic SSC process in Table VIII, i.e., one SSC group separately for each different handbag model. Then after style-based SSC recognition, the identified SSC is the final recognized handbag model. Therefore, no color-based handbag model recognition is needed. There are two ways for the automatic SSC recognition: 1) style-based SSC recognition without color information, and 2) style-based SSC recognition with color information. For the first way, we just need to repeat the experiment for style-based SSC recognition and recognize handbag models. For the second way, we concatenate three types of features: Type-I patch feature g, Type-II patch feature g' and the color feature, where the color feature is extracted from K patches via color naming method. It can be seen from the Table VIII that an automatic SSC grouping procedure with color information can achieve better performance than other compared frameworks for handbag recognition.



(a) (b)

Fig. 13. Training and testing time based on different number of SSCs: (a) overall training time (min.) and (b) testing time for each query handbag ROI (sec.).

C. Computational Complexity

We report the time complexity of training and testing for the proposed method (taking BrandBag-I as an example). Experiment is conducted on Matlab R2012b, in a desktop of Intel(R) Core(TM) i5-4570 3.20 GHz CPU, and 32.0GB RAM. Fig. 13(a) gives the curve for measuring the total training time (in min.) changing with the increasing number of SSCs. Fig. 13(b) plots the testing time (in sec.) for each query.

For Type-II patch selection, we adopt all the C_N^2 binary classifiers to keep the measurement fair and simple. However, it is not necessary to train all the binary classifiers and the speed would become much faster if we reduce the number of binary classifiers to be trained. When $K = 125$, and we do not adopt any binary classifier, the training time will reduce to 16.05 seconds. This will sacrifice approximately 2% in accuracy. Moreover, we compute the overall training and testing time sequentially for Type-I, Type-II patch selection and color element selection modules. To speed up, these modules can be computed in parallel. Based on Fig. 13(b), our handbag search engine is able to be applied for real-time.

VI. CONCLUSION

In this paper, we present a novel method to recognize handbag models. Our study focuses on the discriminative representations for handbags. For style-based SSC recognition, two different mid-level handbag patch selection algorithms are proposed to find the discriminative handbag patches. Then a low-level complementary feature is designed to capture the fine details of the handbag patches. For color-based handbag model recognition, the dominant color naming features are selected and used to represent the color of the handbag. The experimental results show that our method performs very well on recognizing handbags. In our future work, more elaborate patch partition methods are needed to deal with the non-rigid deformation of handbags. More compact color descriptor specialized in expressing handbags can be designed. It will also be desired to develop a large scale branded handbag dataset.

ACKNOWLEDGMENT

This work was carried out at the Rapid-Rich Object Search (ROSE) Laboratory at the Nanyang Technological University, Singapore. The ROSE Laboratory was supported by the National Research Foundation, Singapore, under its Interactive Digital Media Strategic Research Programme.

REFERENCES

[1] W. Yin, J. Luo, and C. W. Chen, "Event-based semantic image adaptation for user-centric mobile display devices," *IEEE Trans. Multimedia*, vol. 13, no. 3, pp. 432–442, Jun. 2011.

[2] G. Gualdi, A. Prati, and R. Cucchiara, "Video streaming for mobile video surveillance," *IEEE Trans. Multimedia.*, vol. 10, no. 6, pp. 1142–1154, Oct. 2008.

[3] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Fast image retrieval: Query pruning and early termination," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 648–659, May 2015.

[4] Y.-H. Kuo, W.-H. Cheng, H.-T. Lin, and W. Hsu, "Unsupervised semantic feature discovery for image object retrieval and tag refinement," *IEE Trans. Multimedia*, vol. 14, no. 4, pp. 1079–1090, Aug. 2012.

[5] Q. You, J. Yuan, J. Wang, P. Guo, and J. Luo, "Snap n' shop: Visual search-based mobile shopping made a breeze by machine and crowd intelligence," in *Proc. IEEE Int. Conf. Semantic Comput*, Feb. 2015, pp. 173–180.

[6] Y. Liu and T. Mei, "Optimizing visual search reranking via pairwise learning," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 280–291, Apr. 2011.

[7] R. Ji *et al.*, "Learning to distribute vocabulary indexing for scalable visual search," *IEEE Trans. Multimedia*, vol. 15, no. 1, pp. 153–166, Jan. 2013.

[8] P. Li, M. Wang, J. Cheng, C. Xu, and H. Lu, "Spectral hashing with semantically consistent graph for image indexing," *IEEE Trans. Multimedia*, vol. 15, no. 1, pp. 141–152, Jan. 2013.

[9] S. Liu *et al.*, "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3330–3337.

[10] S. Liu *et al.*, "Hi, magic closet, tell me what to wear!" in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 619–628.

[11] L. Liu *et al.*, "wow! you are so beautiful today!," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 11, no. 1s, pp. 20:1–20:22, 2014.

[12] S. Liu *et al.*, "Fashion parsing with weak color-category labels," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 253–265, Jan. 2014.

[13] T. Chen, K.-H. Yap, and D. Zhang, "Discriminative soft bag-of-visual phrase for mobile landmark recognition," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 612–622, Apr. 2014.

[14] B. Yao, G. Bradski, and L. Fei-Fei, "A codebook-free and annotation-free approach for fine-grained image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3466–3473.

[15] O. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3498–3505.

[16] S. Branson *et al.*, "Visual recognition with humans in the loop," in *Proc. 11th Eur. Conf. Comput. Vis., Part IV*, 2010, pp. 438–451.

[17] B. Yao, A. Khosla, and L. Fei-Fei, "Combining randomization and discrimination for fine-grained image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 1577–1584.

[18] T. Berg and P. Belhumeur, "Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 955–962.

[19] J. Deng, J. Krause, and L. Fei-Fei, "Fine-grained crowdsourcing for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 580–587.

[20] K. Duan, D. Parikh, D. Crandall, and K. Grauman, "Discovering localized attributes for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3474–3481.

[21] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 729–736.

[22] A. Angelova and S. Zhu, "Efficient object detection and segmentation for fine-grained recognition," presented at the *IEEE Conf. Comput. Vis. Pattern Recog.*, Portlant, OR, USA, pp. 23–28, Jun. 2013.

[23] S. Yang, L. Bo, J. Wang, and L. G. Shapiro, "Unsupervised template learning for fine-grained object recognition," in *Proc. Adv. Neural Inf. Process. Syst. 25*, 2012, pp. 3131–3139.

[24] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 1, pp. 886–893.

[26] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[27] J. Wang *et al.*, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 3360–3367.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.

[29] R. Farrell *et al.*, "Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 161–168.

[30] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200–2011 dataset," California Instit. of Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.

[31] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Grap. Image Process.*, Dec. 2008, pp. 722–729.

[32] A. Angelova, S. Zhu, and Y. Lin, "Image segmentation for large-scale subcategory flower recognition," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Jan. 2013, pp. 39–45.

[33] E. Gavves, B. Fernando, C. Snoek, A. Smeulders, and T. Tuytelaars, "Local alignments for fine-grained categorization," *Int. J. Comput. Vis.*, vol. 111, no. 2, pp. 191–212, 2015.

[34] E. Gavves, B. Fernando, C. G. M. Snoek, A. W. M. Smeulders, and T. Tuytelaars, "Fine-grained categorization by alignments," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1713–1720.

[35] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 923–930.

[36] I. Endres, K. J. Shih, J. Jiaa, and D. Hoiem, "Learning collections of part models for object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 939–946.

[37] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 73–86.

[38] P. Over *et al.* "Trecvid 2015—An overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proc. TRECVID*, 2015.

[39] H. Li, E. Kim, X. Huang, and L. He, "Object matching with a locally affine-invariant constraint," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 1641–1648.

[40] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2911–2918.

[41] H. Li, J. Huang, S. Zhang, and X. Huang, "Optimal object matching via convexification and composition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 33–40.

[42] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[43] J. Lim, C. Zitnick, and P. Dollar, "Sketch tokens: A learned mid-level representation for contour and object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 3158–3165.

[44] J. Shotton *et al.*, "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 1297–1304.

[45] J. Shotton *et al.*, "Efficient human pose estimation from single depth images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2821–2840, Dec. 2013.

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

[47] Y. Wang, S. Li, and A. C. Kot, "Complementary feature extraction for branded handbag recognition," in *Proc. 12th IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 5896–5900.

[48] T. Stojic, I. Reljin, and B. Reljin, "Adaptation of multifractal analysis to segmentation of microcalcifications in digital mammograms," *Physica A: Stat. Mech. Appl.*, vol. 367, pp. 494–508, 2006.

[49] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," *Comput. Vis. Image Underst.*, vol. 106, pp. 59–70, 2007.

[50] G. Griffin, A. Holub, and P. Perona, "Caltech-256 Object Category Dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2007-001, 2007.

[51] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 413–420.

[52] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 2, pp. 2169–2178.

[53] L.-J. Li and F.-F. Li, "What, where and who? Classifying events by scene and object recognition," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[54] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei, "Novel dataset for fine-grained image categorization," presented at the *1st Workshop Fine-Grained Vis. Categorization, IEEE Conf. Comput. Vis. Pattern Recog.*, Colorado Springs, CO, USA, Jun. 2011.

[55] P. Welinder *et al.*, "Caltech-UCSD Birds 200," California Instit. of Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2010-001, 2010.

[56] J. Lim, C. Zitnick, and P. Dollar, "Sketch tokens: A learned mid-level representation for contour and object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 3158–3165.

[57] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun, "Salient object detection by composition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1028–1035.

[58] A. Veldaldi and A. Zisserman, "Recognition of object instances practical," 2016. [Online]. Available: http://www.robots.ox.ac.uk/vgg/practicals/instance-recognition/index.html

[59] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 1794–1801.

[60] A. Abdel-Hakim and A. Farag, "CSIFT: A SIFT descriptor with color invariant characteristics," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 2, pp. 1978–1983.

[61] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1512–1523, Jul. 2009.

[62] F. Shahbaz Khan *et al.*, "Color attributes for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3306–3313.

[63] L. Jia Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & Semantic feature sparsification," in *Proc. Adv. Neural Inf. Process. Syst. 23*, 2010, pp. 1378–1386.

[64] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," *CoRR*, 2013. [Online]. Available: http://arxiv.org/abs/1310.1531

Yan Wang (S'13) received the B.S. degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2011, and is currently working toward the Ph.D. degree in electrical and electronic engineering at the Nanyang Technological University, Singapore.

She was an exchange student with the Tokyo Institute of Technology International Research Opportunities Program, Tokyo, Japan, in 2012. Currently, she is a Research Scholar with the Computational Biology and Cognitive Science Laboratory, Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA. Her current research interests include computer vision, object recognition, machine learning, and cognitive science.

Sheng Li received the Ph.D. degree in electrical and electronic engineering from the Nanyang Technological University, Singapore, in 2013.

From 2013 to 2015, he was a Research Fellow with the Rapid Rich Object Search Laboratory, Nanyang Technological University, Singapore. He is currently with the School of Communication and Information Engineering, Shanghai University, Shanghai, China. His research interests include biometric template protection, pattern recognition, multimedia forensics, and security.

Dr. Li was the recipient of received the IEEE WIFS Best Student Paper Silver Award.

Alex C. Kot (S'85-M'89-SM'98-F'06) has been with the Nanyang Technological University, Singapore, since 1991. He headed the Division of Information Engineering at the School of Electrical and Electronic Engineering for eight years and served as an Associate Chair (Research) and Vice Dean Research for the School of Electrical and Electronic Engineering. He is currently a Professor and Associate Dean for the College of Engineering, the Director of Rapid-Rich Object Search (ROSE) Laboratory, and the Director of the NTU-PKU Joint Research Institute. He has authored or coauthored works in the areas of signal processing for communication, biometrics, data-hiding, image forensics, information security, and image object retrieval and recognition.

Dr. Kot is a Fellow of IES and a Fellow of the Academy of Engineering, Singapore. He served as the Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE SIGNAL PROCESSING LETTERS, the *IEEE Signal Processing Magazine*, the IEEE JOURNAL OF THE SPECIAL TOPICS IN SIGNAL PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON INFORMATION, FORENSICS, AND SECURITY, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS PART II and PART I. He has served the IEEE Signal Processing Society in various capacities, such as the General Co-Chair for the 2004 IEEE International Conference on Image Processing and the Vice President for the IEEE Signal Processing Society. He was the recipient of the Best Teacher of the Year Award and is a co-author for several Best Paper Awards including ICPR, IEEE WIFS, and IWDW. He is the IEEE Distinguished Lecturer for the Signal Processing Society and the Circuits and Systems Society.