IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 2, 2020, accepted March 14, 2020, date of publication March 23, 2020, date of current version April 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2982560

Large Scale Category-Structured Image Retrieval for Object Identification Through Supervised Learning of CNN and SURF-Based Matching

XIAOQING LI, JIANSHENG YANG, AND JINWEN MA
Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing 100871, China
Corresponding author: Jinwen Ma (jwma@math.pku.edu.cn)

This work was supported by the Natural Science Foundation of China under Grant U1604153.

ABSTRACT In the modern era of Internet, mobile and digital information technology, image retrieval for object identification, just as wine label retrieval from a wine bottle image, has become an important and urgent problem in artificial intelligence. In comparison with the general image retrieval, it is rather challenging because there are a huge number of object identification or brand images which are very similar and difficult to discriminate, and the number of different brand images in the given dataset changes greatly, that is, the samples are strongly unbalanced for these brands. In this paper, we propose a CNN-SURF Consecutive Filtering and Matching (CSCFM) framework for this kind of image retrieval, specifically focalizing on wine label retrieval. In particular, Convolutional Neural Network (CNN) is utilized to filter out the impossible main-brands (manufacturers) for narrowing down the range of retrieval and the Speeded Up Robust Features (SURF) matching is improved by adopting the RANdom SAmple Consensus (RANSAC) mechanism and the modified Term Frequency–Inverse Document Frequency (TF-IDF) distance for the accurate retrieval of the sub-brand (item attribute under the manufacture). The experiments are conducted on a dataset containing approximately 548k images of wine labels with 17, 328 main-brands and 260, 579 sub-brands. It is demonstrated by the experimental results that our proposed method can solve the wine label retrieval problem effectively and efficiently. Moreover, our proposed method is further evaluated on two pubic benchmarks of the object identification image retrieval tasks, Oxford Buildings Benchmark (Oxford5k) and the University of Kentucky of Indoor Things Benchmark (UKB), and achieves 88.3% mean average precision and 3.92 N-S score in Oxford5k and UKB, respectively.

INDEX TERMS Image retrieval, object identification, wine label retrieval, CNN, SURF descriptor, filter out the impossible main-brands.

I. INTRODUCTION

Image retrieval has been one of classical research fields in computer vision and image processing. Nowadays, with the development of information technology and database, Image retrieval technology has brought great convenience to our work and life. Face retrieval [1] can help polices and other security personnels catch suspects more quickly. In online shopping, commodity images retrieval [2] can help customers find their favorite commodities. Building retrieval [3] from the map can help people locate themselves more accurately and reduce the possibility of getting lost. Clothes retrieval [4] can help buyers find out the clothes they want. Bird retrieval [5] can help people know the type of the bird and learn more about the nature. In fact. this is essentially a problem of object identification through image retrieval. That is, there is a database of the images which contains the same or similar objects with different categories or identities and we need to detect or recognize the object identity containing in each image. With the development of Internet, mobile and digital information technology, the image retrieval for object identification becomes an important and urgent problem in artificial intelligence. However, in many practical applications, it is rather challenging because there are a huge number of images which are very similar and difficult to discriminate, and the samples are strongly unbalanced for these categories.

The associate editor coordinating the review of this manuscript and approving it for publication was Long Wang.

We investigate the object identification image retrieval from a specific task: wine label retrieval. In our daily life, wine is one of the most widely consumed beverages in the world and thus automatic wine label retrieval becomes very important and popular. In a wine label retrieval system, as a user inputs a wine label image taken by a mobile phone or camera, the retrieval result can help the user get the wine brand and other related information he would like to know. In this way, it helps the user to recognize and learn about the wine more easily. However, there are hundreds of thousands of red wines in the world and we cannot set up a complete database of all the standard wine label images. So, the wine label retrieval system begins from a small database of wine label images and collects more samples of wine label images from the customers step by step. After a long time accumulation, we can have a large database with a huge number of wine label images, which are generally distributed unbalancedly on the wine labels. Our main task is to establish an effective and efficient wine label retrieval system for such a database.

We then turn to the existing methods of image retrieval. In fact, they have been developed into two main streams: conventional and Convolutional Neural Network (CNN) based methods. Conventional image retrievals are based on an aggregation of conventional image features by the methods such as Bag of Words (BOW) [6], Vector of Locally Aggregated Descriptors (VLAD) [7] and Fisher Vector [8]. In fact, the color, shape, texture, and Scale-Invariant Feature Transform (SIFT) [9] features are all typical conventional features. Following the increased interest of deep neural networks, CNN based image retrieval approaches become active in recent years. These approaches can reduce the semantic gap in the task of image retrieval in comparison with conventional retrieval methods. Their features are extracted from fully connected layers [10] or convolutional layers [11], and then employed to match with the methods such as SVM or softmax regression. In addition to extracting features from the whole image, CNN models also extract the features from local regions of the image. For example, Regional Maximum Activation of Convolutions (RMAC) [12] and Multiscale Regional Maximum Activation of Convolutions (MS-RMAC) [13] are classical image retrieval approaches. Those methods may be effective in general image retrieval problems, but they are not feasible for wine label retrieval because there are three major challenges on wine label datasets. Firstly, there is a huge number of wine label images with large numbers of main-brands and sub-brands, which makes the retrieval task more complicated. If we apply a conventional method to solve wine label retrieval, both time and space costs will be high. Secondly, the numbers of samples for different brands are quite different, and the numbers of samples for different sub-brands are also quite different, with certain brands having only one sample. We can see the real case from Table 4 in our experiment. This inter-class unbalance can lead to poor performance even if CNN based methods are used for this task. Thirdly, there is

(a) (b)

FIGURE 1. Some examples of the same wine label brand. (a) and (b) are two pairs of examples with the same main-brand and the same sub-brand, respectively, but the images in (b) have a more significant difference than those in (a).

(a) (b)

FIGURE 2. Some examples of different wine label brands. (a) and (b) are two pairs of examples with different main-brands and different sub-brands, but the images in (b) have a more slight difference than those in (a).

even a significant difference among some wine label images of the same sub-brand, while there is a slight difference among some wine label images of different brands, as shown in Figure1 and Figure2, which will make the retrieval more difficult.

As for wine label retrieval, one naive way is to get the wine label region by using the edge-based method, and then implement fuzzy c-means clustering on its individual wine letters to recognize the text [14]. While this system can achieve a high recognition accuracy in some special wine label images, it has some obvious disadvantages. First of all, it works well only if the text on the wine label is English. However, there always exist some texts on the wine labels that are not English. Then, it is well known that the letter recognition in this system relies heavily on the detection of candidate text region. If the fonts of letters are standard in wine label images, the detection performs well, so does the retrieval. Clearly, the realistic situation is that the fonts in wine label images are usually changeable which can lead to poor detection. Another available way is to use a hierarchical feature and a client-sever searching architecture proposed by Wu *et al.* [15]. In fact, it was built by the Speeded Up Robust Features (SURF) descriptors [16], K-D tree and k-means method. Although it can maintain a high recognition accuracy, its retrieval time cost for each image will increase greatly on a large database.

In order to overcome these difficulties, we recently proposed the framework of CNN-SIFT Consecutive Searching and Matching (CSCSM) [17] for the task of wine label retrieval with a large database. It is a two-phase system with consecutive searching and matching. In this paper, we further extend our previous study of the CSCSM framework methodologically and theoretically. Specifically, we design the framework with a new version of CNN architecture and

the SURF descriptor, which is referred to as the CNN-SURF Consecutive Filtering and Matching (CSCFM) framework. In fact, it is effective for the large scale image retrieval with a dataset which has a category structure, i.e., whose samples can be naturally classified into a number of categories (corresponding to main-brands). So, this is a kind of category-structured image retrieval. In such a situation, the samples in the dataset can be classified into different categories and the number of categories is relatively stable as compared with that of objects which are described by the images. It should be noted that most of image retrieval tasks are essentially category-structured. For example, the samples for commodity image retrieval can be classified into clothes, fruits, computers and so on. The main colors of birds in the task of birds image retrieval can be served as its categories. Moreover, the division of categories in certain image retrieval tasks is not necessary to be very clear, and an imprecise or even fuzzy category division can be remarkable to reduce the computational time with our CSCFM framework. In addition, the number of samples in a dataset of category-structured image retrieval can be increased as we need in practice while the final retrieval performance of our CSCFM framework will not be influenced evidently, because our CSCFM approach does not need a very accurate CNN classifier and the downstream improved SURF matching is not impacted by this total sample number. Therefore, our CSCFM framework work in this paper can be easily applied to a variety of large scale category-structured image retrieval tasks. As for the task of wine label retrieval, all wine images can be categorized into different manufacturers, and then CNN is utilized to recognize the possible coarse-grained main-brands (manufacturers), i.e., to filter out the impossible main-brands and narrow down the range of retrieval. On the other hand, the SURF matching is improved by adopting the RANdom SAmple Consensus (RANSAC) mechanism and the modified Term Frequency-Inverse Document Frequency (TF-IDF) distance for the accurate retrieval of the fine-grained sub-brand (item attribute under the manufacture). We conduct extensive experiments on a dataset containing approximately 548k wine images with 17, 328 main-brands and 260, 579 sub-brands. It is demonstrated by the experimental results that our proposed CSCFM method can solve the wine label retrieval problem effectively and efficiently. We also evaluate the CSCFM method on two pubic object identification image retrieval benchmarks, Oxford Buidings Benchmark(Oxford5k) and the University of Kentucky Benchmark (UKB) and achieve new state-of-the-art retrieval indexes of 88.3% mean average precision and 3.92 N-S score in the two cases, respectively.

The main contributions of our work are as follows:

- A two-phase image retrieval framework is designed and exploited for category-structured object identification, particularly for wine label image retrieval. During the coarse-to-fine retrieval process, it can retrieve the main-brand and sub-brand of an inputted wine label image.
- The combination of deep CNN architecture and the improved SURF descriptor is proposed to reduce the semantic gap in the image retrieval without losing local information.
- We improve the SURF matching by adopting the RANSAC and our modified TF-IDF distance that can reduce the computational cost greatly. In fact, this modified TF-IDF distance does not need to build the conventional codebook and new word embedding, whose training time is rather expensive.
- It is demonstrated by the experiments on two additional public benchmarks that our proposed CSCFM framework outperforms the state-of-the-art methods on category-structured object identification image retrieval.

The rest of this paper is organized as follows. Section II presents an overview of related works. The CSCFM framework is introduced in Section III. Section IV presents the related experimental results and comparisons, followed by concluding remarks in Section V.

II. RELATED WORK

For wine label or general object identification image retrieval, we need to implement the Fully Convolutional Network (FCN) to locate and cut off the accurate label region so that the retrieval process can be much more effective. We further need the CNN and SURF descriptor to establish our CSCFM framework. In this section, we introduce these models and related researches.

A. FCN

The architecture of FCN was firstly proposed by Long *et al.* [18] to segment an image in a semantic mode. They adapted advanced classification networks into fully convolutional networks and transfered their learned representations by fine-tuning to the segmentation task. They then designed a novel skip architecture which combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentations. In fact, FCN turns the pixel-level classification to the semantic segmentation. Instead of using the full connection layer to get a fixed length feature vector for different classification, it can process images in any size and make a pixel-level prediction by utilizing the up-sampling layer to recover the images size. FCN has achieved good performance on many segmentation tasks. In our framework, we apply FCN to separate the wine label or bottle region from the background area.

B. CNN

CNN was firstly designed for the recognition of English letters in 1989. Krizhevsky *et al.* [19] developed it to AlexNet that had achieved the lowest classification error rate on the large scale image dataset ImageNet at that time. Then, another version of CNN, Visual Geometry Group Network (VGGNet), was further developed by Simonyan and Zisserman [20] with much more convolution layers than AlexNet. To overcome the vanishing and exploding of gradient in

deeper layers of neural networks, He *et al.* [21] proposed Residual Network (ResNet) by adopting the Residual strategy to further improve the classification accuracy. After that, Chen *et al.* [22] constructed Dual Path Network (DPN) which has superior performances on classification accuracy, training speed and model size.

In addition to image classification, CNN has also applied to image retrieval. Wan *et al.* [23] indicated that the neuron activation of CNN can be served as generic features in image retrieval. Gong *et al.* [24] combined VLAD with CNN activation from deep convolution layers to form the complete features. According to the performance of these CNN based retrieval methods, we can see that CNN features can effectively reduce the semantic gap and have great potential for image retrieval. However, the success of these methods depends on whether there are enough data to train their networks. As for the wine label dataset, this premise cannot be guaranteed. In fact, lots of main-brands have only 10 labeled samples and the samples under each main-brand also have different sub-brands. Besides, some sub-brands have a few examples. As a result, the sample sizes of both main-brands and sub-brands are insufficient to train a reasonable CNN. Considering these situations, our proposed CSCFM framework uses a CNN to reduce retrieval range rather than extract features for the retrieval of an input image. This CNN is not used to return one accurate main-brand, but a limited number of main-brands (It will be introduced in Section III-B in detail), i.e. the CNN classifier of our proposed CSCFM framework does not have to be very accurate. We then make the SURF matching of the input image to the images under these main-brands and return the final retrieval results. The actual experiments demonstrated that our strategy can get much higher accuracy with less retrieval time.

C. SURF

SURF descriptor is one of the most popular interest point descriptors. Bay *et al.* [16] firstly introduced the SURF algorithm as a scale and rotation invariant interest point detector and descriptor. In fact, it is an improved version of SIFT algorithm. The major difference between SIFT and SURF is on the implementation of scale-space. SIFT implements the image pyramid where the input image is iteratively convolved with Gaussian kernel and repeatedly sub-sampled [25], while SURF creates the scale-space by applying a group of kernels where the scale increases to the consistent value of the original image. Some studies [26]–[28] have shown that SURF outperforms SIFT in terms of retrieval result and computational time, especially for the images with brightness or blur variation. Actually, in our wine label database, the changes of brightness and ambiguity are widespread, so we choose SURF instead of SIFT as the method of feature extraction for fine-grained matching in the CSCFM framework.

SURF descriptor has been used for image retrieval in certain studies. Mumar [29] and Velmurugan and Baboo [30] all proposed an image retrieval system based on the SURF algorithm for feature detection. Mukherjee *et al.* [31] proposed a novel image retrieval scheme using random forest-based semantic similarity measures and SURF-based bag of visual words and demonstrated its superior performance. However, the above studies are only effective for small datasets. In fact, the feature representation capability of SURF is limited because there exists semantic gap in image retrieval. Besides, the length of the codebook is difficult to determine and the training time is particularly long. Moreover, both the storage and computing costs are enormous on a large dataset. To get rid of these problems, our work is firstly to use a trained CNN to shrink retrieval range and then to implement the SURE matching on a limited number of images so that the storage and computing costs can be reduced greatly. On the other hand, the SURE matching can be improved by adopting the RANSAC and TF-IDF mechanisms. In fact, the RANSAC algorithm [32] can effectively remove the wrong SURF matching pairs by estimating the parameters, while Term Frequency-Inverse Document Frequency (TF-IDF) [33] is a weighting technique for information retrieval and data mining. Specifically, TF measures the frequency of the occurrence where a word appears in a document, and IDF measures the importance of the word on distinguishing different documents. In order to reduce the training time and space storage, we do not use the Bag-Of-Words (BOW) strategy as commonly used in the TF-IDF scheme [34], and modify the TF-IDF distance into a simplified form which can be computed directly. In this way, our improved SURF matching becomes more accurate with the help of our modified TF-IDF distance and the RANSAC mechanism.

III. OUR APPROACH

In this section, we present the overall description of our proposed CSCFM framework and the detailed descriptions of its components.

A. THE OVERALL DESCRIPTION OF THE CSCFM FRAMEWORK

Due to the limited storage space and computing ability of mobile devices, we construct our CSCFM framework with a client-server architecture for wine label image retrieval and its overall flow is shown in Figure 3. In the client site, the user can capture a wine label image using the camera of mobile device and send it as an input query to the server site via the wireless network for wine label retrieval. In the server site, the FCN is firstly utilized to segment the wine label area and remove the background on this input image. Then, the fine-tuned DPN network based on the given labeled dataset returns a number of possible main-brands of the image. Our improved SURF matching is further made on all wine label samples with these possible main-brands to get the most similar images. Therefore, the retrieval result about the main-brand and sub-brand is obtained from the original information description of the similar images. Finally, the server site transmits the retrieval result to the user in the client site.

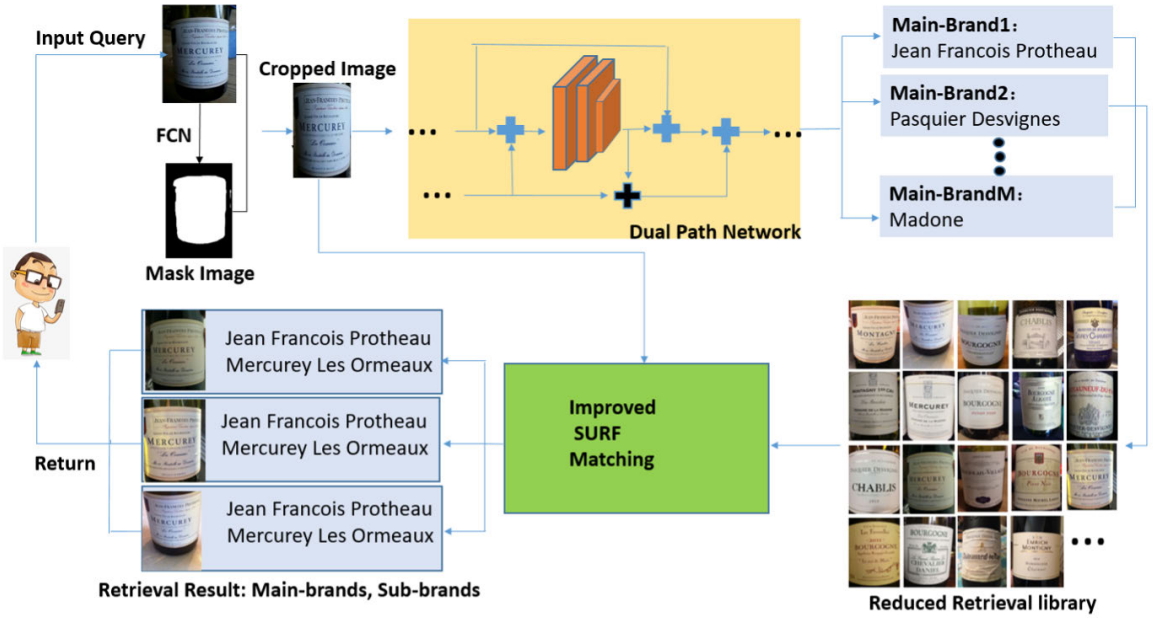


FIGURE 3. The overall flow of the CSCFM framework.

B. FCN BASED SEGMENTATION OF WINE LABEL

In order to reduce the interference of background factors in a wine label image and effectively focus on the wine label, we need to separate the wine label or bottle region from the background area. This is essentially a binary semantic image segmentation problem. As FCN has achieved the good perfprmance in segmentation, we apply FCN to separate the wine label or bottle region from the background area in a supervised learning mode. The wine label image segmentation procedure is shown in Figure 4. With the strategy of transfer learning, we firstly design the FCN with VGGNet-16 and pre-trained on the VOC2012 dataset, and then finetune on our labeled wine label dataset. Specifically, this dataset is collected by ourselves for wine label segmentation. It contains about 9,000 wine label images and each of them includes the wine label area and the background area being marked artificially. By implementing the trained FCN on each wine label image, we get the mask of the wine label region in the image, and then segment the wine label region from the image and remove the background area according to the mask.

C. CNN BASED POSSIBLE MAIN-BRAND RECOGNITION

In order to improve retrieval efficiency, we train a CNN to narrow down the range of retrieval by recognizing a goup of possible main brands of the wine in an image. It is known from the literature and experimental experiences that the ResNet [21] can reuse the features implicitly by using residual connections, but it is not so good at exploring new features; on the contrary, the densely connected network [35] can explore new features continuously, but may cause the feature redundancy; DPN inherits both the advantages of Residual Network (ResNet) and Dense Convolutional Network (DenseNet). In addition, as compared to the other deep

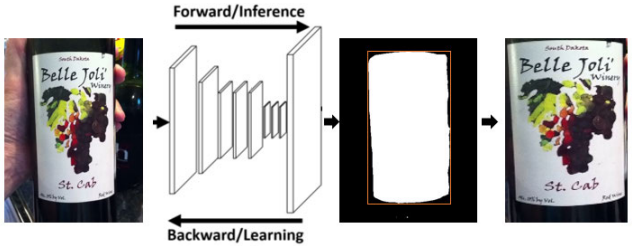


FIGURE 4. The process of using the FCN to segment the wine label region and remove the background area.

CNNs, DPN has higher accuracy, faster training speed and smaller model size. For getting more powerful features to reduce the semantic gap in the retrieval task and saving the computing cost, we utilize DPN to recognize the possible main-brands of a wine label image from a large wine label dataset. Because there exist severely insufficient samples and enormous number of main brand categories in the dataset, it is still difficult to guarantee the classification accuracy of DPN unless a transfer learning strategy is adopted. To reduce the risk of error, we use the multiple possible solutions instead of the best one like [36], that is, to return multiple possible main-brands at the same time. In the light of these ideas, we have the following adaptive strategy.

Let the probability that the query image x_i belongs to main-brand j be p_j^i, where $1 \leq j \leq Z$ and Z is the number of main-brands. N is the number of images in the retrieval dataset. Without loss of generality, let $p_1^i \geq p_2^i \geq \cdots \geq p_Z^i$. In order to ensure the accuracy of retrieval, we set the number of returned main-brands by

$$M_1^i = argmin\{z | \sum_{j=1}^{z} p_j^i \geq \delta\} \tag{1}$$

where $z = 1, 2, \ldots, Z$ and δ is the threshold value set by experience. Generally, we can have a reasonable number of returned main-brands as δ is selected reasonably. However, in certain cases, for example, x_1 and x_2 represent two query samples with $p_1^1 = p_2^1 = \cdots = p_Z^1$ and $\delta/2 \approx p_1^2 > p_2^2 = \cdots = p_Z^2 \approx 0$, respectively. If δ is large, both M_1^1 and M_1^2 are too large. In this case, too many main-brands will be returned. In other words, DPN can not play a role in reducing the search range which will increase the difficulty in SURF matching later. Conversely, if the threshold δ is small, M_1^i is also small as we want. But, the probability that the ground-truth of x_1 is in the returned main-brand set is $p^1 = M_1^1/N$, which will be very small and severely affect the accuracy of the method. So, δ is quite important and should be carefully selected. It has been demonstratated by the experiments that the value of δ is negatively affected by the accuracy of CNN and the average sample size of per main-brand. That is, if the CNN classifier is accurate enough or there are plenty of samples for each main-brand, $\delta \in [0.6, 0.95)$ is good enough to generate a reasonable reduced retrieval library to accelerate the downstream feature matching procedure. Conversely, a relatively larger $\delta \in [0.95, 1)$ is a better choice to make up for the lack of accuracy of CNN or extend the feature matching range to cover more candidate samples.

To get rid of too large number of returned main-brands in the extreme situations, we set an upper bound of the number of main-brands to be a finite number M_2^i for the query image x_i, which is also determined by experience. Finally, the number M of main-brands returned by DPN can be determined by the minimum of M_1^i and M_2^i. That is,

$$M = min\{M_1^i, M_2^i\} \tag{2}$$

In a similar way as the setting of δ, a smaller M_2^i should be selected in the situation of a more accurate CNN or larger average sample size of per main-brand, and vice versa.

D. IMPROVED SURF MATCHING FOR THE FINAL SUB-BRANDS

After getting the most possible main-brands of the query image with the DPN, we implement the SURF matching procedure to retrieve the final sub-brands. Moreover, we modify the TF-IDF distance in such a specific situation and combine the RANSAC and modified TF-IDF mechanisms into the SURF matching procedure to further improve the retrieval accuracy so that an improved SURF matching mechanism is proposed for the final sub-brands of the CSCFM system.

SURF is based on the Hessian Matrix, and uses the Difference-of-Gaussian (DoG) as a basic accurate approximation of Hessian determinant. Then, it uses a distribution of Haar-wavelet responses over all the points of the image to extract the interest points or descriptors with the Haar-wavelet responses being greater than a certain threshold value. With the advanced architecture, SURF algorithm is a speedup and improved version of of SIFT algorithm with the robustness of scale and rotation. Furthermore, for the images with brightness and blur variation, the SURF algorithm has a better matching effect than the SIFT algorithm. Actually, in our wine label image database, the changes of brightness and ambiguity are widespread. Thus, we choose SURF instead of SIFT as the component of our feature extraction.

For the accurate match between two images with the SURF features, it is essential to correctly match the corresponding interest points. However, there usually exist some false matching pairs, especially when the images are as complicated as the wine label images. In order to overcome this difficulty, we adopt the RANSAC algorithm into the SURF matching as done in [32] so that we can successfully delete some mismatching point pairs, and improve the matching performance, as shown in Figure5.

It is clear that the SURF descriptors or interest points of an image contribute differently on image retrieval. If the weights or importance levels of the descriptors are adopted in the matching distance, the image retrieval can be more accurate and effective. In fact, the weights can be estimated under the TF-IDF framework by considering the SURE descriptors as visual words, which thereby needs to build a proper codebook as well as a word embedding in a common way under the BOW strategy. However, the time cost of training a codebook and a new word embedding cannot be undertaken because the training dataset is so huge in our wine label image retrieval. In order to overcome this difficulty, we modify the TF-IDF distance into a simplified form which can be computed directly from certain introduced functions. In this way, we also adopt the modified TF-IDF distance into the SURF matching so that our improved SURF matching becomes more accurate and effective.

Firstly, we formally describe our modified TF-IDF distance. Let x_0 be a query image and $\{x_1, x_2, \ldots, x_N\}$ be the image dataset. For each image, we extract the first K SURF descriptors in the descending order of Haar-wavelet response. $\{s_1^i, s_2^i, \ldots, s_K^i\}$ denote the SURF descriptors of x_i. Let d_{pk}^j be the Euclidean distance between the $p-th$ SURF descriptor of x_0 and the $k-th$ SURF descriptor of x_j. Without loss of generality, let $d_{p1}^j < d_{p2}^j < \cdots < d_{pK}^j$. For the image retrieval, if the SURF descriptor s_p^0 is noise, such as salt and pepper noise, usually $d_{p1}^j/d_{p2}^j \ll 1$; if s_p^0 is the real feature point of x_0, such as contour folding, then $d_{p1}^j/d_{p2}^j \approx 1$. Given the facts above, we define the matching result in s_p^0 and s_k^j as an indicator function by

$$\theta_{pk}^j = \begin{cases} 1, & k \in G_p^j \\ 0, & k \notin G_p^j \end{cases} \tag{3}$$

where

$$G_p^j = \{k | d_{pk}^j = \min_l \{d_{pl}^j\}, d_{pr}^j = \min_{l \neq k} \{d_{pl}^j\}, \ d_{pk}^j/d_{pr}^j \geq \epsilon\}$$

and $\epsilon \in (0, 1]$ is the threshold parameter. Denote

$$TF_p = \frac{\phi(s_p^0)}{\sum_k \phi(s_k^0)} \tag{4}$$

FIGURE 5. The left and right figures illustrate the SURF matching results without and with the RANSAC mechanism, respectively, which show that the SURF matching with the RANSAC mechanism can effectively remove some mismatching pairs.

where $\phi(\cdot)$ is an indicator function and $\phi(s_k^0) = 1$ if $s_k^0 \in \{s_1^0, s_2^0, \ldots, s_K^0\}$, otherwise $\phi(s_k^0) = 0$.

Obviously, TF_p represents the frequency that the p_{th} SURF descriptor of x_0 appears in the image. Since each SURF descriptor of the image is unique, $TF_p = 1/K$.

As for the query image x_0, the IDF of the p_{th} SURF descriptor is denoted as $IDF_p = N/c_p$, where c_p indicates the number of images containing s_p^0. In order to avoid a real number divided by zero, let $c_p = 1$ if there is not eligible image. As a result, we define: $c_p = max\{|\{j|s_p^0 \in \{s_1^j, s_2^j, \ldots, \ldots, s_K^j\}\}|, 1\}$. So, we have the weighted SURF-based distance between x_0 and x_j as follows.

$$d_j = \sum_{p=1}^{K} TF_p IDF_p \sum_{k=1}^{K} \theta_{pk}^j \tag{5}$$

For convenience of the calculation, we equivalently use the following simplified version instead of d_j

$$D_j = \sum_{p=1}^{K} \frac{1}{c_p} \sum_{k=1}^{K} \theta_{pk}^j \tag{6}$$

as our modified TF-IDF distance in this special situation. As there is no need to build a codebook and a new word embedding, the computation and storage of this modified TF-IDF distance are relatively small. As demonstrated in our following experimental results, the computation of the modified TF-IDF distance is almost the same as that of the conventional SURF-based distance, while the retrieval performance with the former is much better than that with the latter.

We then use the modified TF-IDF distance to measure the degree of difference between a query image x_0 and any image in the image dataset or retrieval library. In fact, it is very beneficial to match the true wine images with the same main-brand and sub-brand. The compositional complexity of our modified TF-IDF distance is $O(K^2H^2)$, where H is the dimension of the SURF descriptor. So, it is still polynomial time as same as that of common Euclidean distance. The details of this improved SURF matching algorithm is given in Algorithm 1.

For example, as shown in Table 1, x_0 represents the query image, A, B, C, D are four images in the retrieval library. $s_1^0, s_2^0, s_3^0, s_4^0, s_5^0$ are point pairs matched between x_0 and A, B, C, D after eliminating the mismatching by RANSAC. The number "1" indicates that this SURF descriptor is matched, while the number "0" indicates that it is not matched. "Raw Score" represents the SURF matching score function without TF-IDF, and "Improved Score" represents the SURF matching score function with the TF-IDF mechanism. The conventional SURF matching algorithm calculates the score function only according to the number of SURF descriptors matched, ignoring the importance of individual descriptor. The result of the conventional SURF matching is that A is the best matching to x_0. our improved SURF matching considers the difference in the importance of each SURF descriptor, for example, s_5^0 is a unique feature between x_0 and D, it is more important than the other SURF descriptors in the matching process, and our decision is that D is the best matching to x_0.

Algorithm 1 The Improved SURF Matching Algorithm

Require: query image x_0, retrieval library $\{x_1, \ldots, x_N\}$
Ensure: S best matching images

1: set $i = 1$
2: Extract the SURF descriptors of x_0 and x_i, then implement the SURF matching between x_0 and x_i, and further improve the matching results with RANSAC, as suggested in [32]. The first K SURF descriptors of x_i are extracted and denoted as $\{s_1^i, s_2^i, \ldots, s_K^i\}$.
3: Calculate θ_{pk}^i by Eq.(3), where $p, k = 1, 2, \ldots, K$.
4: Calculate the parameter c_p by

$$c_p = max\{|\{j|s_p^0 \in \{s_1^j, s_2^j, \ldots, \ldots, s_K^j\}\}|, 1\}$$

5: Calculate the modified TF-IDF distance D_i by

$$D_j = \sum_{p=1}^{K} \frac{1}{c_p} \sum_{k=1}^{K} \theta_{pk}^j$$

6: If $i < N$, then $i = i + 1$ and go to step 2, otherwise go to the next step.
7: Sort $D_1, D_2, \ldots, D_N$ in the descending order.
8: Take the first S distances and return the matching images.

TABLE 1. An example of the improved SURF matching result.

Retrieval Library	SURF s_1^0	s_2^0	s_3^0	s_4^0	s_5^0	Raw Score	Improved Score
A	1	1	1	1	0	**4**	17/12
B	1	1	0	1	0	3	11/12
C	1	0	1	1	0	3	13/12
D	1	1	0	0	1	3	**19/12**

IV. EXPERIMENTAL RESULTS

In this section, we test and compare the CSCFM framework by carrying out the following experiments:

- The improved SURF matching and comparison with the conventional SURF and SIFT matching.
- CSCFM retrieval and comparison on a general wine label image dataset to demonstrate the effectiveness of the proposed framework.

TABLE 2. Retrieval accuracies and times of the improved SURF matching method and competitive methods on our filtered wine label image dataset.

Algorithm	MA[a]	SA[b]	Time[c] (s)				
			FCN Segmentation	Feature Extraction	RANSAC	Feature Matching	Total
SIFT	0.58	0.48	0.6031	0.6771	-	16.0260	17.3062
SIFT+RANSAC	0.74	0.69	0.5967	0.6224	5.0867	16.1026	22.4084
SIFT+RANSAC+TF-IDF	0.82	0.76	0.5975	0.6290	5.0416	17.7519[d]	24.0200
SURF	0.58	0.50	0.6016	0.4757	-	8.8478	**9.9251**
SURF+RANSAC	0.74	0.70	**0.5910**	0.4412	4.0298	**8.7792**	13.8412
SURF+RANSAC+TF-IDF	**0.83**	**0.78**	0.6003	**0.4347**	4.0730	9.0553[d]	14.1633

[a]The average retrieval accuracy of the main-brands.
[b]The accuracy of the sub-brands.
[c]The average time spent for the retrieval of one image.
[d]The feature matching using the modified TF-IDF distance.

- CSCFM retrieval and comparison on a large real world wine label image dataset.
- CSCFM retrieval and comparison on the Oxford buildings dataset.
- CSCFM retrieval and comparison on the University of Kentucky Benchmark dataset.

Without specification, we make all these experiments based on Python in Ubuntu 16.04. The hardware configuration is as follows: NVIDIA Tesla M40 graphics card, 24GB GPU memory, E5-2620 v4 @ 2.10GHz s × 2 processor. The CNNs used in the experiments are all pre-trained on ImageNet [19]. In the finetune process of each CNN, we adopt the synchronized SGD training in a single GPU, starting with a learning rate of 0.01. In order to accelerate the training process and reduce the overfitting, we train the network with the Batch Normalization (BN) [39] enabled. We utilize a momentum of 0.9 and a weight decay of 10^{-4}. The dropout [40] of 0.5 is used before the final classifier layer.

A. THE IMPROVED SURF MATCHING PERFORMANCE AND COMPARISON

We firstly evaluate the retrieval effect of the improved SURF matching method and compare it with the conventional SURF matching on a manually filtered wine label image dataset. There are 500 images on this dataset, including 100 query images which represent 100 kinds of wine and a retrieval library of 400 images. In the retrieval library, 100 images are under the same main-brands and sub-brands for 100 query images, 100 images only have same main-brands but different sub-brands for 100 query images, and the remaining 200 images have completely different main-brands for query images. An example of the dataset images are illustrated in Figure 6. For clarity, the improved SURF matching method is referred to as SURF+RANSAC+TF-IDF, while the SURF matching combined with the RANSAC mechanism is referred to as SURF+RANSAC. SIFT related algorithms are named in the same way. In order to reduce the influence of noise and other interference factors, we extract only the first 500 SURF keypoints according to the degree of Haar-wavelet responses. The experimental results are listed in Table 2.

It can be seen from Table 2 that the SURF descriptors are more suitable for wine label retrieval than the SIFT descriptors. Moreover, our improved SURF matching method leveraged with the RANSAC mechanism and the modified TF-IDF distance is effective and efficient for wine label retrieval. The retrieval accuracy of the conventional SURF matching method becomes lower when the retrieval range is larger. The SURF+RANSAC matching method introduces the RANSAC process so that the retrieval time is increased in comparison with the pure SURF matching. But the MA is increased by 27.6% and the SA is increased by 40.0%. In comparison with SURF+RANSAC, SURF+RANSAC+TF-IDF increases the MA by 12.2% and the SA by 11.4%, with the retrieval time being increased slightly by 2.3%, which demonstrates the efficiency and effectiveness of our modified TF-IDF distance. Moreover, the comparison between SIFT+RANSAC and SIFT+RANSAC+TF-IDF also leads to the same conclusion. As a result, our modified TF-IDF distance has a big potentiality to enhance the general image matching methods.

FIGURE 6. Typical wine label images in the retrieval library.

TABLE 3. Retrieval accuracies and times of CSCFM and competitive methods on the general wine label image dataset, being detailed with the FCN segmentation time T_S, CNN classification time T_C, feature extraction time T_{FP}, feature matching time T_{TM} and the total retrieval time T, respectively.

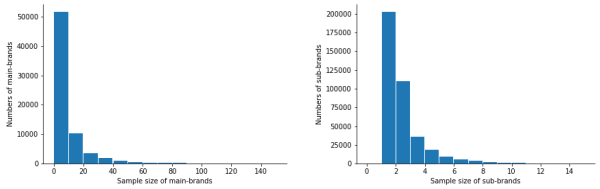
Feature Type	Method	MA[a]	SA[b]	Time[c](s)					
				T_S	T_C	T_{FE}	T_{FP}	T_{FM}	T
SURF Based	Improved SURF Matching	0.8867	0.7096	-	-	0.5551	-	402.6612	403.0163
	BOW(800)	0.3374	0.1972	-	-	**0.2413**	0.2831	0.0573	**0.5818**
	VLAD(100)	0.6981	0.5465	-	-	0.3021	0.2343	0.0762	0.6126
CNN Based	CNNfeat-SVM1	0.9167	-	0.5254	-	0.6971	-	**0.0271**	1.2496
	CNNfeat-SVM2	0.7976	0.1905	0.5038	-	0.6743	-	4.5150	5.6931
Fusion	OR [55]	0.9031	0.7339	-	-	1.3287	0.4815	0.3913	2.2015
	CSCSM	**0.9285**	0.7976	0.4187	0.6143	0.4306	-	2.8330	4.2096
	CSCFM	**0.9285**	**0.8277**	0.4325	0.5091	0.3005	-	1.8032	2.8452

[a]The average retrieval accuracy of the main-brands.
[b]The accuracy of the sub-brands.
[c]The average time spent for the retrieval of one image.

B. CSCFM RETRIEVAL AND COMPARISON ON A GENERAL WINE LABEL IMAGE DATASET

We further implement the CSCFM method on a general wine label image dataset, and compare it with the competitive algorithms including our improved SURF matching method, the BOW and VLAD algorithms with our improved SURF descriptors, the CNNfeat-SVM algorithm proposed in [37], and the Object-level Representation (OR) algorithm proposed in [54]. Because the BOW and VLAD algorithms adopt the K-means algorithm to establish the codebook, their time complexity is $O(NKT)$, where N, K, T denote the number of training samples, the codebook length and the times of iteration, respectively. Taking into account that if the algorithms are implemented on a large dataset, the cost of time will be very high. So, we begin to take a subset of the large wine label dataset as a general size experimental dataset. In fact, we randomly select 13,929 wine label images which belong to 115 main-brands and 6,582 sub-brands. For test, we randomly select 115 query images. For fairness, the CNN models for comparison are ResNeXt-50 [38] which are pre-trained on ImageNet as well as our large wine label dataset which does not contain 115 query images, and then finetuned on this general dataset. For the BOW and VLAD algorithms, we set the length of the codebook to be 800 and 100, respectively. In order to speed up the retrieval process, we use the Ball-Tree algorithm for the codebooks of BOW and VLAD algorithms, which can reduce the time complexity of each image retrieval from O(N) to O(logN). For the CNNfeat-SVM algorithm, we extract the CNN features from the last fully connected layer of ResNeXt-50 and use SVM for classification. The multiple classification strategy is 1-VS-1 in SVM. In order to compare the ability of CNN fully-connected layer features to express the main-brands and sub-brands, we conduct two experiments on the CNN-SVM algorithm, namely CNNfeat-SVM1 and CNNfeat-SVM2. The image labels for training can be set to the main-brand or sub-brand for the different purpose. In the finetune process of ResNeXt-50, the batch size is 100. As for the OR algorithm, the BING detector [54] is firstly implemented to detect potential objects, and then the SIFT features are extracted and aggregated by the VLAD algorithm to form an image representation, while the CNN features are extracted to form another image representation, and finally the two representations are fused together to be indexed into an inverted table and encoded with the Product Quantization (PQ) scheme for the image retrieval. The CSCSM framework in [17] also has a two-phase structure as the CSCFM framework, but utilizes the ResNeXt for learning the main-brands. In order to reduce the interference of background factors in wine label images, all the algorithms firstly use the identically trained FCN to segment the wine label region. The experimental results of all the above algorithms are listed in Table 3.

From the experimental results in Table 3, we can see that the improved SURF matching method has a high recognition accuracy for the wine label sub-brand, but the time cost is also high. In comparison with the improved SURF matching method, the computational times of BOW and VLAD significantly reduce, but their retrieval accuracies are rather low. CNNfeat-SVM1 has a recognition accuracy of 91.67% for the main-brands. It benefits from that the dataset has the relatively less number of main-brands and the computation time of the algorithm does not depend on the retrieval library size through the CNN. The algorithm uses only 1.2496 seconds per a query image. However, it cannot be used to identify the sub-brands, which makes it unusable to our wine label retrieval. Compared with CNNfeat-SVM1, CNNfeat-SVM2 can identify the sub-brands, but the retrieval accuracy of the main-brands is remarkably reduced. CNN-based methods outperform the SURF-based methods, which demonstrates the effectiveness of semantic-aware CNN features. However, SIFT may be superior to CNN if the images are rather confusing in the semantic space. In fact, SIFT describes low-level details, while CNN describes general semantics. It seems that none of them is systematically better than the other. So it is a good choice to adopt both features to achieve a better result. The OR algorithm fuses the CNN and SIFT features. However, in the CNN finetune phase, the sample sizes of different sub-brands are small. Moreover, there is even a remarkable difference among some wine label images of the same sub-brand, while there is only a slight difference among some wine label images of different brands. As a result, the discrimination of CNN features cannot be very



(a) (b)

FIGURE 7. (a) The real-world large wine label image dataset has different sample sizes for the main-brands; (b) The same as (a) except for the sub-brands instead of the main-brands.

TABLE 4. The numbers of samples for the main-brands and sub-brands,respectively.

Number of images	Number of main-brands
$11 \sim 20$	8974
$21 \sim 30$	3359
$31 \sim 50$	2811
$51 \sim 100$	1473
$101 \sim 1371$	711
Number of images	**Number of sub-brands**
1	129932
2	71257
3	24993
$4 \sim 10$	32802
$11 \sim 371$	1595

strong. Unlike the OR algorithm, the CNN classifiers of our CSCSM and CSCFM frameworks do not have to be very accurate, and they are just used to filter out some irrelevant images. The down-stream feature matching procedure acts as a makeup for CNN classifiers. Benefited from this strategy, our CSCSM and CSCFM methods both achieve the best accuracy on the recognition of main-brands by reaching 92.85%. Because the improved SURF descriptors are more suitable for wine label retrieval than the improved SIFT descriptors, our CSCFM method has the higher recognition accuracy for sub-brands and even takes less time. These experimental results on this typical wine label dataset demonstrate that our proposed CSCFM method outperforms other competitive methods on wine label retrieval.

C. CSCFM RETRIEVAL AND COMPARISON ON THE LARGE WINE LABEL IMAGE DATASET

We now implement the CSCFM method on a real-world large wine label image dataset which was provided by Ruixun Science and Technology (Beijing) Limited Company in China. It contains 17,328 main-brands, 260,579 sub-brands and total 547,857 wine images. The examples of the wine label images in the dataset are already illustrated in Figure 6. The above two datasets in the experiments are actually the subsets of this dataset. The relevant information of the main-brands and sub-brands on this dataset is given in Table 4 and Figure7.

All of the images in this large dataset come from manual shooting using mobile phones. The size of the images is 500×375 and all the images are formatted into RGB. It contains many interference factors, such as background, fingers, light changes, local highlight, marginal highlight, image rotation and so on. Each image is labeled with a main-brand and a

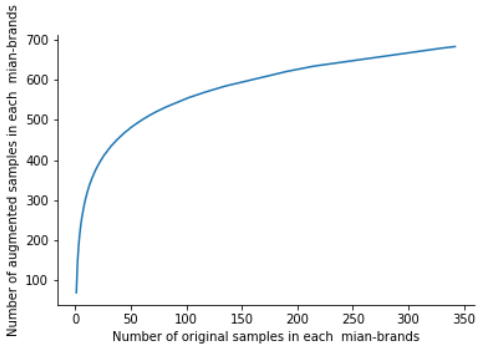


FIGURE 8. The result of the data enhancement, where the horizontal axis indicates the number of samples for each main-brand, and the vertical axis indicates the number of samples for these main-bands after data enhancement.

sub-brand. In our experiment, the ratio of the training and test samples is set to 4:1. In order to reduce the imbalance of training data, we firstly perform the data enhancement including adding Gaussian blur, changing contrast, sharpness, saturation, brightness, and tilt. Figure8 shows the sample sizes of the enhanced main-brands. After that, the FCN is implemented to segment the wine label region and remove the background area so that we can reduce the interference from the background.

We then implement DPN92 [22] to narrow down the range of retrieval by recognizing a number of possible main-brands. So, we need to train the DPN92 in the supervised mode with the training data. In the finetune process of DPN92, the batch size is 40. Finally, we use the improved SURF matching to get the final retrieval result.

Because the number of images, the number of main-brands and the number of sub-brands for the dataset are all large, most of the methods we mentioned above in Subsection IV-B can not work for wine label retrieval on this dataset. For the pure improved SURF matching, the retrieval time is unbearable which force us to give up its experiment on this large dataset. As for the SURF matching using the BOW and VLAD strategy, while the retrieval time may be acceptable, the training time of the codebooks is very long. The CNNfeat-SVM1 can not be used for this task, because it is incompetent in retrieving sub-brands. As for the CNNfeat-SVM2, the accuracy of sub-brands retrieval will less than that in Table 3, because each sub-brand has the average of only 2.10 examples in the wine label image dataset. It is far from satisfied for the training process. For the other advanced image retrieval methods, due to their unpublished source codes or intolerable time cost in this large scale wine label dataset, we will only compare them with our proposed method according to the results of their papers on the public datasets in the next two subsections. Therefore, we mainly implement the CSCFM method on this real-world wine label dataset. During the experiment, we set δ and M_2 to 95% and 5 to control the retrieval range of SURF matching. At the same time, we also compare CSCFM with two other methods. One is the CSCSM framework with ResNet101 and the

TABLE 5. Retrieval accuracies and times of CSCFM and competitive methods on the large wine label image dataset.

Method	MA[a]	SA[b]	Time[c](s)
CSCSM	0.9107	0.784	9.5657
DPN+ISIFT	0.9201	0.8232	7.8981
CSCFM	**0.9223**	**0.8294**	**2.8956**

[a]The average retrieval accuracy of the main-brands.
[b]The accuracy of the sub-brands.
[c]The average time spent when each image is retrieved.

improved SIFT matching. The other is the CSCSM framework with DPN92 and the improved SIFT matching, being referred to as DPN+ISIFT. Their experimental results are listed in Table 5. From these results, we can find out that the CSCFM method proposed in this paper is more effective, which can get the better retrieval result with less time on the large real-world wine label dataset. As shown in Table 3, searching among 13929 images costs about 2.8452 seconds per query image. While in Table 5, searching among 547,857 images costs about only 2.8956 seconds per query image, which takes just 0.0504 second more than Experiment IV-B. The reason is that in FCN segmentation and CNN filtering procedure, time spent in both is very small and not relative to the scale of retrieval datasets. The major time cost is in our improved SURF matching. It depends on the size of the reduced retrieval dataset after CNN filtering. However the size of the reduced retrieval dataset is usually very small. So that the time cost of the down-stream improved SURF matching will not be impacted obviously by the scale of total retrieval dataset. That is, different from common image retrieval methods, our method can overcome the problem of time boosting as the retrieval datasets going larger.

D. CSCFM RETRIEVAL AND COMPARISON ON OXFORD BUILDINGS DATASET

We continue to evaluate the CSCFM method on a similar benchmark of category-structured object identification image retrieval, Oxford Buildings dataset [41]. Actually, the Oxford Buildings dataset contains 5062 images downloaded from Flickr by searching for particular landmarks. The collection is manually annotated to generate a comprehensive ground truth for 11 different landmarks, each represented by 5 possible queries. This gives a set of 55 queries for evaluation. The mean average precision (mAP) is used to measure the retrieval performance over the 55 queries. Some typical images of the dataset are shown in Figure 9.

All the images except the 55 query images in the Oxford Buildings dataset are taken as the training dataset. In order to use a classification neural network to narrow down the range of retrieval, we divide all the images into 11 categories according to the number of different buildings in the images. We firstly perform the training data enhancement including adding Gaussian blur, changing contrast, sharpness, saturation, brightness, and tilt so that the classification neural network can learn the features of Oxford building images more robustly. We then utilize ResNet-18 [21] to learn the

FIGURE 9. Typical images of the Oxford buildings dataset (Oxford 5k).

main-brands. Actually, our ResNet-18 classifier is pretrained on the ImagesNet and finetuned only on the training set of Oxford Buildings dataset. So, we do not use any extra prior information in comparison with the other methods. In the finetune process of ResNet-18, the batch size is 100. In the retrieval process, we set δ and M_2 to 95% and 5, respectively, to reduce the retrieval range. Finally, we use the improved SURF matching to get the final retrieval result.

We compare the CSCFM method with state-of-the-art methods for the Oxford Buildings dataset in the literature. Table 6 gives the experimental results of the CSCFM and state-of-the-art methods on the Oxford Buildings dataset. It should be noted that some methods like the attentive deep local features based method proposed by Noh *et al.* [42], enhance their results by finetuning the feature extractor on a special landmarks dataset which is large-scale and similar to the Oxford Buildings dataset. However, our CSCFM method pretrains its CNN classifier only on the ImageNet. This is unfair to our CSCFM method. Even so, as shown in Table 6, our CSCFM method still achieves the best result among all the approaches. If we use the large-scale landmarks dataset to finetune our CNN classifier, we believe that our CSCFM method can achieve a higher mAP. Moreover, as compared with the other conventional feature and CNN feature fused methods, our proposed CSCFM method achieves a significant improvement of mAP over 4.9%. The key reason may be that the use of CNN in our CSCFM framework is quite different from those of the other fusion methods. In our approach, CNN is utilized as a classifier to shrink the down-stream retrieval range, and does not need to be very accurate. While it is used as an appended feature to the conventional features directly for the image matching, the retrieval accuracy is influenced obviously by the quality of CNN features. The experimental

TABLE 6. Retrieval accuracies of the CSCFM and state-of-the-art methods on the Oxford buildings dataset.

Feature Type	Method	mAP
Conventional	Jegou et al. [47]	56.0
	Arandjelovic [55]	80.9
CNN Based	Seddati et al. [44]	72.3
	Kalantidis et al. [46]	68.2
	Li et al. [13]	68.9
	Tolias et al. [12]	66.9
	Radenovic et al. [45]	79.7
	Noh et al. [42]	83.8
	Gordo et al. [43]	86.1
Fusion	Zhang et al [53]	81.6
	Zheng et al. [52]	83.4
	CSCFM (Ours)	**88.3**

FIGURE 10. Typical examples of the images in UKB.

results show that our proposed strategy is evidently better for this task than those of Zhang *et al.* [53] and Zheng *et al.* [52].

E. CSCFM RETRIEVAL AND COMPARISON ON THE UNIVERSITY OF KENTUCKY BENCHMARK

We finally evaluate the CSCFM approach on another public retrieval benchmark of category-structured object identification image retrieval, the University of Kentucky Benchmark (UKB) [48]. In fact, UKB contains 2,550 classes which each class has 4 images in JPEG format, and total 10,200 images. The images of this dataset are mostly taken indoors, such as plants, toys, clocks and CD covers. They are taken with different perspective, illumination and distance conditions. All the images have a resolution of 640 $\times$ 480 pixels. Some typical examples of the images in the dataset are shown in Figure 10. We use the standard evaluation metric, N-S score, to evaluate the performance of image retrieval. That is, for each group, we select one image as the query and compute the recall at its four retrieval results.

In our CSCFM method, all the images except the 2550 query images in UKB are taken as the training dataset for the CNN. It is observed that the shooting angles of images in the same label are very different. So, our data enhancement especially emphasizes the rotation enhancement. We also add Gaussian blur, change contrast, do sharpness and change brightness in the data enhancement. We utilize ResNeXt-50 finetuned on our large wine label dataset to learn the features from the training data. In the finetune process of ResNeXt-50, the batch size is 60. In the retrieval process, we set δ and M_2 to 95% and 1 to shrink the retrieval range. In the test phase, this CNN can shrink the retrieval range by recognizing the query image. Finally, we use the improved SURF matching to get the final retrieval result.

TABLE 7. Retrieval accuracies of the CSCFM and state-of-the-art methods on UKB.

Feature Type	Method	N-S Score
Conventional	Jegou et al. [47]	3.53
	Zheng et al. [51]	3.62
	Qin et al. [50]	3.67
	Zheng et al. [49]	3.85
CNN Based	Gordo et al. [43]	3.84
	Seddati et al. [44]	3.91
Fusion	Zheng et al. [11]	3.84
	Sun et al. [54]	3.81
	Zheng et al. [52]	3.88
	CSCFM (Ours)	**3.92**

We also compare our CSCFM method with the state-of-the-art methods for UKB in the literature. Table 7 contains the experimental results of the CSCFM method and state-of-the-art methods on UKB.

As shown in Table 7, our CSCFM method achieves 3.92 N-S score, being the best result among all the methods. This again shows that our proposed CSCFM framework has the ability to be extended to the general category-structured image retrieval tasks for object identification.

V. CONCLUSION

We have established the CNN-SURF Consecutive Filtering and Matching (CSCFM) framework for the large scale category-structured image retrieval of object identification, especially for wine label image retrieval. This framework is a two-phase system of consecutive filtering and matching. Specifically, it utilizes the CNN trained for the main-brand classification to narrow down the range of retrieval or filter out the impossible main-brands of a query image, and then modify the TF-IDF distance and adopt the RANSAC mechanism and this modified TF-IDF distance into the SURF descriptors to construct the improved SURF matching to get the final sub-brand of the query image. This strategy makes the retrieval time complexity independent on the number of samples for the retrieval dataset. It is demonstrated by the experiments on a real-world wine label image dataset that the CSCFM method can retrieve both the main-brands and sub-brands of wine label effectively and efficiently. It is further demonstrated by the experiments that the CSCFM method has the ability to solve the general category-structured image retrieval problems of of object identification like the Oxford Buildings Benchmark (Oxford5k) and the University of Kentucky Benchmark (UKB) and achieve the competitive accuracy in comparison with the state-of-the-art retrieval methods. Clearly, our CSCFM framework using the combination of CNN classification and accurate feature matching in the two phases can be easily utilized to many

other category-structured image retrieval tasks. Moreover, our modified TF-IDF distance gives a more effective and efficient way to apply the TF-IDF mechanism to large scale image retrieval. This strategy, we believe, can leverage the performances of various feature matching methods, not limited to SURF matching and SIFT matching.

REFERENCES

[1] S. R. Dubey, "Local directional relation pattern for unconstrained and robust face retrieval," *Multimedia Tools Appl.*, vol. 78, no. 19, pp. 28063–28088, Oct. 2019.
[2] Y. Zhang, P. Pan, Y. Zheng, K. Zhao, Y. Zhang, X. Ren, and R. Jin, "Visual search at Alibaba," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 993–1001.
[3] A. Jimenez, J. M. Alvarez, and X. Giro-i-Nieto, "Class-weighted convolutional features for visual instance search," 2017, *arXiv:1707.02581*. [Online]. Available: http://arxiv.org/abs/1707.02581
[4] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5337–5345.
[5] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked CNN for fine-grained visual categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1173–1182.
[6] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, p. 1470.
[7] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3304–3311.
[8] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier, "Large-scale image retrieval with compressed Fisher vectors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3384–3391.
[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[10] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki, "Visualinstance retrieval with deep convolutional networks," *ITE Trans. Media Technol. Appl.*, vol. 4, no. 3, pp. 251–258, 2016.
[11] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian, "Query-adaptive late fusion for image search and person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1741–1750.
[12] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," 2015, *arXiv:1511.05879*. [Online]. Available: http://arxiv.org/abs/1511.05879
[13] Y. Li, Y. Xu, J. Wang, Z. Miao, and Y. Zhang, "MS-RMAC: Multiscale regional maximum activation of convolutions for image retrieval," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 609–613, May 2017.
[14] J. Lim, S. Kim, J. Park, G. Lee, H. Yang, and C. Lee, "Recognition of text in wine label images," in *Proc. Chin. Conf. Pattern Recognit.*, Nov. 2009, pp. 1–5.
[15] M.-Y. Wu, J.-H. Lee, and S.-W. Kuo, "A hierarchical feature search method for wine label image recognition," in *Proc. 38th Int. Conf. Telecommun. Signal Process. (TSP)*, Jul. 2015, pp. 568–572.
[16] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.*, May 2006, pp. 404–417.
[17] X. Li, J. Yang, and J. Ma, "CNN-SIFT consecutive searching and matching for wine label retrieval," in *Proc. Int. Conf. Intell. Comput.*, 2019, pp. 250–261.
[18] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representat*, May 2015, pp. 1–14.
[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[22] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4467–4475.
[23] J. Wan, D. Wang, S. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: A comprehensive study," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 157–166.
[24] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 392–407.
[25] C. Evans, "Notes on the opensurf library," Univ. Bristol, Bristol, U.K., Tech. Rep. CSTR-09-001, Jan. 2009.
[26] S. He, C. Zhang, and P. Hao, "Comparative study of features for fingerprint indexing," in *Proc. 16th IEEE Int. Conf. Image Process. (ICIP)*, Nov. 2009, pp. 2749–2752.
[27] M. Dawood, C. Cappelle, M. E. El Najjar, M. Khalil, and D. Pomorski, "Harris, SIFT and SURF features comparison for vehicle localization based on virtual 3D model and camera," in *Proc. 3rd Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Oct. 2012, pp. 307–312.
[28] L. Juan and G. Oubong, "SURF applied in panorama image stitching," in *Proc. 2nd Int. Conf. Image Process. Theory, Tools Appl.*, Jul. 2010, pp. 495–499.
[29] A. Mumar, "Image retrieval using SURF features," M.S. thesis, Thapar Univ., Patiala, India, 2011.
[30] K. Velmurugan and S. S. Baboo, "Content-based image retrieval using SURF and colour moments," *Global J. Comput. Sci. Technol.*, vol. 11, no. 10, pp. 1–5, May 2011.
[31] A. Mukherjee, J. Sil, and A. S. Chowdhury, "Image retrieval using random forest-based semantic similarity measures and SURF-based visual words," in *Proc. 2nd Int. Conf. Comput. Vis. Image Process.*, Singapore, 2018, pp. 79–90.
[32] Y. Chen, Q. Sun, H. Xu, and L. Geng, "Matching method of remote sensing images based on SURF algorithm and RANSAC algorithm," *J. Frontiers Comput. Sci. Technol.*, vol. 6, no. 9, pp. 822–828, 2012.
[33] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. Eur. Conf. Mach. Learn.*, Chemnitz, Germany, New York, NY, USA: Springer, Apr. 1998, pp. 137–142.
[34] Y. Kalantidis, G. Tolias, E. Spyrou, P. Mylonas, Y. Avrithis, and S. Kollias, "VIRaL: Visual image retrieval and localization," *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 555–592, 2011.
[35] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[36] K. Tieu and P. Viola, "Boosting image retrieval: Special issue on content-based image retrieval," *Int. J. Comput. Vis.*, vol. 56, pp. 17–36, Jan. 2004.
[37] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "Factors of transferability for a generic ConvNet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1790–1802, Sep. 2016.
[38] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.
[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167
[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[41] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[42] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3456–3465.
[43] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *Int. J. Comput. Vis.*, vol. 124, no. 2, pp. 237–254, Sep. 2017.
[44] O. Seddati, S. Dupont, S. Mahmoudi, and M. Parian, "Towards good practices for image retrieval based on CNN features," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1246–1255.
[45] F. Radenovic, G. Tolias, and O. Chum, "CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands, New York, NY, USA: Springer, Oct. 2016, pp. 3–20.
[46] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands, New York, NY, USA: Springer, Oct. 2016, pp. 685–701.

[47] H. Jegou and A. Zisserman, "Triangulation embedding and democratic aggregation for image search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2014, pp. 3310–3317.
[48] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 2161–2168.
[49] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Packing and padding: Coupled multi-index for accurate image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1939–1946.
[50] D. Qin, C. Wengert, and L. Van Gool, "Query adaptive similarity for large scale object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1610–1617.
[51] L. Zheng, S. Wang, W. Zhou, and Q. Tian, "Bayes merging of multiple vocabularies for scalable image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1955–1962.
[52] L. Zheng, S. Wang, J. Wang, and Q. Tian, "Accurate image search with multi-scale contextual evidences," *Int. J. Comput. Vis.*, vol. 120, no. 1, pp. 1–13, Oct. 2016.
[53] G. Zhang, Z. Zeng, S. Zhang, Y. Zhang, and W. Wu, "SIFT matching with CNN evidences for particular object retrieval," *Neurocomputing*, vol. 238, pp. 399–409, May 2017.
[54] S. Sun, W. Zhou, Q. Tian, and H. Li, "Scalable object retrieval with compact image representation from generic object regions," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 12, no. 2, pp. 1–21, Oct. 2015.
[55] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2911–2918.

XIAOQING LI received the B.S. degree in mathematics from the Ocean University of China, Qingdao, China, in 2016. She is currently pursuing the Ph.D. degree in applied mathematics with the School of Mathematical Sciences, Peking University, Beijing, China. Her current research interests include machine learning, image retrieval, and neural networks.

JIANSHENG YANG received the B.S., M.S., and Ph.D. degrees from Peking University, Beijing, China, in 1988, 1991, and 1994, respectively. He is currently a Professor of mathematics with Peking University, where he is also a Faculty Member. His research interests include wavelet analysis, image reconstruction, and computer algorithms.

JINWEN MA received the M.S. degree in applied mathematics from Xi'an Jiaotong University, in 1988, and the Ph.D. degree in probability theory and statistics from Nankai University, in 1992. From July 1992 to November 1999, he was a Lecturer or an Associate Professor with the Department of Mathematics, Shantou University. In December 1999, he became a Full Professor with the Institute of Mathematics, Shantou University. In September 2001, he has joined the Department of Information Science, School of Mathematical Sciences, Peking University, where he is currently a Full Professor and a Ph.D. Tutor. From 1995 to 2003, he also visited several times at the Department of Computer Science and Engineering, The Chinese University of Hong Kong, as a Research Associate or a Fellow. He also worked as a Research Scientist with the Amari Research Unit, RIKEN Brain Science Institute, Japan, from September 2005 to August 2006. He has published over 100 academic articles on neural networks, pattern recognition, bioinformatics, and information theory.