IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 26, 2019, accepted December 19, 2019, date of publication December 25, 2019,
date of current version January 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2962315

Image-Based Service Recommendation System: A JPEG-Coefficient RFs Approach

FARHAN ULLAH[1], BOFENG ZHANG[1], AND REHAN ULLAH KHAN[2,3]
[1]School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
[2]Department of Information Technology, College of Computer, Qassim University, Buraydah 51452, Saudi Arabia
[3]Intelligent Analytics Group (IAG), College of Computer, Qassim University, Buraydah 51452, Saudi Arabia

Corresponding author: Bofeng Zhang (bfzhang@shu.edu.cn)

This work was supported by the National Key R&D Program of China under Grant 2017YFC0907505.

ABSTRACT Online shopping platforms are growing at an unprecedented rate all over the world. These platforms mostly rely on search engines, which are still primarily based on the knowledge-base and use keywords matching for finding similar products. However, customers want an interactive setup that is convenient and reliable for querying related products. In this paper, we propose a novel idea of searching for products in an online shopping system using an image-based approach. A user can provide, select, or click an image, and similar image-based products will be presented to the user. The proposed recommendation system is based on content-based image retrieval and is composed of two major phases; Phase 1 and Phase 2. In Phase 1, the proposed approach learns the class/type of the product. In Phase 2, the proposed recommendation system retrieves closely matched similar products. For Phase 1, the proposed approach creates a model of products using Machine Learning (ML). The model is then used to find the category of the test products. From the ML perspectives, we employ the Random Forests (RF) classifier, and for feature extraction, we use the JPEG coefficients. The dataset used for proof of concepts includes 20 categories of products. For image-based recommendation, the proposed RF model is evaluated for Phase 1 and Phase 2. In Phase 1, the evaluation of the proposed model generates a 75% accurate model. For performance enhancements, the RF model has been integrated into the Deep Learning (DL) setup achieving 84% accurate predictions. Based on the custom evaluation approach for Phase 2, the proposed recommendation approach achieves 98% correct recommendations, thus demonstrating its efficacy for the product recommendation in practical applications.

INDEX TERMS Recommendation system, services, machine learning, random forests, deep learning, SVM.

I. INTRODUCTION

The online retail websites are rapidly evolving and their popularity is exponentially growing. As in the Nielsen Global Connected Commerce Survey (2015) [22], 63% of respondents who bought the products during the past six months stated that they used online services. Online shoppers face the problem of selecting products from a large number of available online products. The number of these products is exponentially increasing due to a large number of companies moving to online businesses. This business growth is useful on the one hand, but on the other hand, it creates a problem for the user to accurately and optimally buy the desired item. Thus these challenges and concerns have raised the demands for recommendation systems to facilitate the users with a convenient and comfortable approach. Almost every E-commerce enterprise nowadays optimize its recommendation system. The optimal recommendation system will not only benefit the customers but will also help the retailer by increasing sales and customer satisfaction.

Traditional E-commerce based search engines are still struggling for an optimal recommendation system because most of these services employ text-based searching. Most of these systems use textual meta-data of products such as attributes, descriptions, and purchase histories of the different users. The textual and key-words based systems can mislead the recommendation system for the inferior quality and non-related products needed by the users. Recently, machine learning algorithms and models have shown its efficacy in detection and recognition scenarios, for example, Neural Networks (NN), Decision Trees (DT), and Deep Learning (DL). Therefore, the traditional search paradigms of the text

The associate editor coordinating the review of this manuscript and approving it for publication was Joey Tianyi Zhou.

description can be replaced or augmented by the visual search for the product recommendation system. A picture of any product should clarify the user's demands for the appearance, usage, and brand of the desired products. While the computer vision and image processing fields have matured, the applications of product retrieval based on the image features using artificial intelligence in the online shopping domain remains mostly unexplored and is open to new findings.

In this paper, based on user interaction and interests, we propose a novel idea to search the products efficiently in an online shopping system using image-based searching techniques. As such, a user provides or selects an image, and similar products (images) are presented to the user. The proposed recommendation system consists of two major phases: In Phase 1, the proposed approach learns the class/type of the product based on the image characteristics. In Phase 2, the proposed recommendation system retrieves closely matched similar products.

For Phase 1, the proposed approach takes advantage of ML for learning the features of the image/product and generates a learned model. The model is then used to find the category/class of query products. Once the category/type of the product is identified, in Phase 2, the JPEG feature vectors-based Euclidian distance is used to retrieve the top 20 matching products from the available items in the particular class of products. These 20 items are then further processed by the proposed "Struct-Hist" approach for retrieving the top 10 most relevant products. The Struct-Hist uses the image feature and is explained in the corresponding section.

From the ML phase in Phase 1 of a product's class learning, we propose the Random Forests (RF) meta-classifier due to its generalization capabilities and excellent performance in state of the art. For learning the class of the products and the feature extraction from images, we use the JPEG coefficients as image features. The dataset used for proof of concept contains product images and labels from the Amazon website. This dataset includes images having 20 categories of products. For image-based recommendation, the proposed RF model is evaluated for Phase 1 and Phase 2. In Phase 1, the evaluation of the proposed model generates a 75% accurate model. For performance enhancements, the RF model is further integrated into the DL setup and achieves 84% accurate predictions. Based on the custom evaluation approach for phase 2, the proposed recommendation approach delivers 98% correct recommendations and demonstrates its efficacy for the recommendation of image-based systems.

From the implementation point of view of the recommendation system, a user submits or selects an image of the product, and similar products (images) are presented to the user. The proposed approach is based on this assumption. However, this assumption also covers the use-cases in which the query image is obtained from the user selection, clicking, and history of user purchases. Thus, our proposed approach is based on learning features from user-based images and recommending similar products based on these images.

The rest of the paper is organized as follows: Section 2 discusses previous related work. Section 3 discusses the ML models and JPEG features. Section 4 explains the proposed approach in the form of Phase 1 and Phase 2. Section 5 evaluates the Phase 1 and Phase 2 of the image-based recommendation approach and presents the comparative analysis. Section 6 concludes the paper.

II. ILITERATURE REVIEW

The authors in [1] present a novel approach for One-Class collaborative filtering that is based on estimating the users' fashion-aware personalized ranking. The recommendation system combines high-level visual features extracted by the Convolutional Neural Network (CNN) from the past feedback and evolving trends. In [2], the proposed approach models the human sense of the relationships between objects based on their appearances. The approach is based on the human perception of visual connections between products. This human perception is modeled as a network inference graph problem. It is thus capable of recommending clothes and accessories together with excellent subjective performance. For image-based recommendation, the authors in [4] aim to recommend images using Tuned Perceptual Retrieval (TPR), Complementary Nearest Neighbor Consensus (CNNC), Gaussian Mixture Models (GMM), Markov Chain (MCL), and Texture Agnostic Retrieval (TAR). The authors report that the CNNC, GMM, TAR, and TPR easy to train; however, CNNC and GMM are complex, while the TPR, GMM, and TAR do not generalize well. In [5], the authors apply the AlexNet [32] based DL that is used to model one thousand different categories of images. The authors in [6] use the CNN model which classifies images into their relevant classes in the ImageNet Challenge. Image similarity has also been investigated in state of the art as in [7] and [8]. However, most of the approaches are based on category similarity, i.e., the products are similar if they are in a similar category. However, there is a possibility that the images might belong to a different category. In such cases, machine learning helps in finding the classes of the images. Therefore, for category retrievals, one approach is to first classify images into their respective categories. Once the category of a particular image is identified, the next step recommends the images from that category. The article [9], uses the NN to calculate the similarities within the category. In [12], the authors focus on learning similarity by using CNN for multi-products in a single image. In [13], the authors concentrate on image similarity and image semantics. The authors in [14] adopt the image descriptions of visual denotations. They propose a new similarity metrics of semantic inference for event descriptions. Similarly, the article [15] proposes the semantic image browser for representing the information visualization with the automated intelligent image analysis. The authors in [23] address the issue of long queries and propose the contextual-based image recommendation system. In [24], the authors use Geo-tagged images from social media for the travel recommendation system. They integrate the time, location, and

weather information with the recorded images. The authors in [25] take advantage of the Geo-tagged images for road-based travel recommendation system. The proposed approach recommends the most famous landmarks and travel routings. The authors use the Spatial Clustering (SC) approach for ranking and identifying the landmarks. In [26], the authors discuss the content-based based approach for image-based recommendation retrieval. The article [27] recommends a personality-based recommendation based on the personality traits in images.

III. CLASSIFICATION AND FEATURES

In this section, we discuss the classification paradigm and the features set used for the recommended approach.

A. RANDOM FOREST (RF)

An RF algorithm by Leo Breiman [17] is a supervised algorithm that generates forests having many trees. In the RF classifier approach, generally, if we have more trees in the forest, the higher is the performance of the classification model. RF model belongs to the ensemble classifier models, which are primarily adapted for problems including classification and regression. The RF algorithm operates by producing decision trees in the training phase. These trees output the class labels, and the final decision of the class depends on the majority voting of the trees. Figure 1 shows the flow of the RF classifier. A sample ''n'' from ''N'' samples is input to the RF classifier. The RF model first generates several trees by using feature subsets. The individual tree presents a classification result. The result of the classification depends on the majority voting system. The test sample is assigned to the class that gets the highest voting scores from the trees.

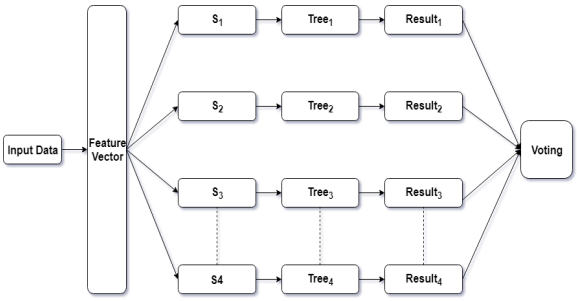

FIGURE 1. A generic representation of RF classifier.

B. CONVOLUTION NEURAL NETWORKS (CNN)

CNN has proved to be a handy and innovative tool of DL to learn image feature sets and inherent relationships in low-level features to higher-level objects in images. The generic architecture of CNN contains interconnected layers. The CNN has repeated convolutional blocks, Rectified Linear Units (ReLU), and pooling layers [32]. Convolutional layers perform the convolution of the input with a set of filters. The filters learned during the training phase. The nonlinearity in the data is modeled by the ReLU layer [32]. The pooling layer samples the input and consolidates the image class.

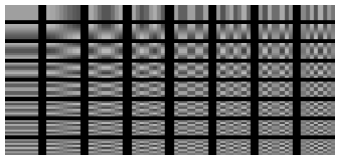

FIGURE 2. The example of 8 × 8 pixel JPEG block matrix.

C. JPEG-COEFFICIENT FEATURES

JPEG Coefficient algorithm is developed by the Joint Photographic Expert Group (JPEG). The JPEG algorithm converts the original input image into the YCbCr color space for compression purposes. The objective is reducing the size of the image while maintaining optimal quality. The image is divided into 8 × 8 or 16 × 16 pixel blocks, as shown in Figure 2. Discrete Cosine Transform (DCT) is applied to the 8 × 8 pixel window which generates 64 values. Then by the JPEG quantization process, the high-frequency values of an image are discarded, and the low-frequency details are preserved. Before the recommendation task, we extract the JPEG image features of each image in our data set. For feature extraction and feature vector construction, we apply the JPEG Coefficient algorithm, which is represented as follows.

$$B_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{mn} cos \frac{\pi\,(2m+1)\,p}{2M} cos \frac{\pi\,(2n+1)\,q}{2N}, \quad 0 \le p \le M-1 \quad 0 \le q \le N-1 \tag{1}$$

where, the parameters α_p *and* α_q are as follows:

$$\alpha_p = \begin{cases} \frac{1}{\sqrt{M}}, & p=0 \\ \sqrt{\frac{2}{M}}, & 1 \le p \le M-1 \end{cases} \qquad \alpha_q = \begin{cases} \frac{1}{\sqrt{N}}, & q=0 \\ \sqrt{\frac{2}{N}}, & 1 \le q \le N-1 \end{cases} \tag{2}$$

Thus the α_p and α_q are normalized scaling factors for orthonormal transformation, P is the horizontal spatial frequency, Q is the vertical spatial frequency. A_{mn} is the pixel value at coordinates (m, n). B_{pq} is the DCT coefficient at coordinates (p, q).

IV. PROPOSED MODEL

From the implementation point of view of the recommendation system, a user submits or selects an image of the product, and similar products/images are presented to the user. This is an assumption of the proposed approach. However, this assumption also covers the use-cases in which the query image is obtained from the user selection, clicking, and history of user purchases. Thus, the proposed approach is based on learning features from user-based images and recommending similar products based on these images. This assumption synchronizes us with the image-based content recognition paradigm of Computer Vision and ML. The proposed recommendation system based on the assumption is shown in Figure 3 and is composed of two major phases:

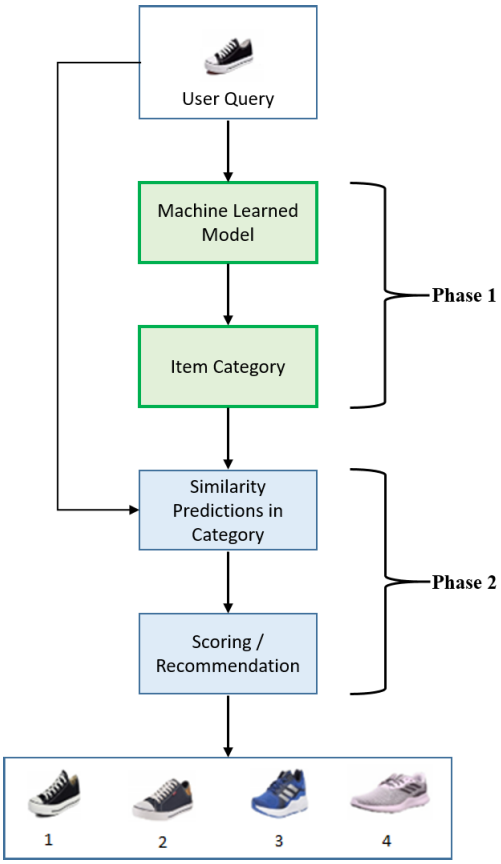

FIGURE 3. The generic flow of the proposed recommendation approach. Green represents "Phase 1" which learns the category of the query image. The blue color represents the "Phase 2" which retrieves similar images from a particular category of images.

Phase 1: Category learning (Green color)
Phase 2: Recommendation (Blue color)

A. "PHASE 1": CATEGORY LEARNING

In Phase 1, the proposed approach models and learns the class/category of the product based on the image characteristics. Once the category is selected by Phase 1, Phase 2 retrieves the closely matched similar products from the corresponding category.

Figure 3 shows the generic flow of the steps for the image-based recommendation system. As the proposed approach is based on the user query image, the approach employs two phases; one for the category/class learning phase and the other for a similar image retrieval phase based on the category. The green color represents the class learning phase. The blue color represents the similar image retrieval from the specific class of images as a recommendation. Phase 2 is further explained in Figure 6.

Figure 4 and Figure 5 show the first phase of the recommendation system, which is the expansion of the green blocks in Figure 3. The first phase is based on the training and testing paradigm of the ML. For the first phase, the approach takes advantage of the ML for learning the features of the image/product characteristics and generates a learned model.

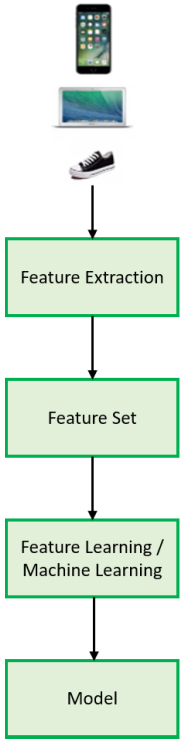

FIGURE 4. The first phase of the recommendation process. The first phase learns the features using the ML classifier to determine the category/class of the products based on the image features.

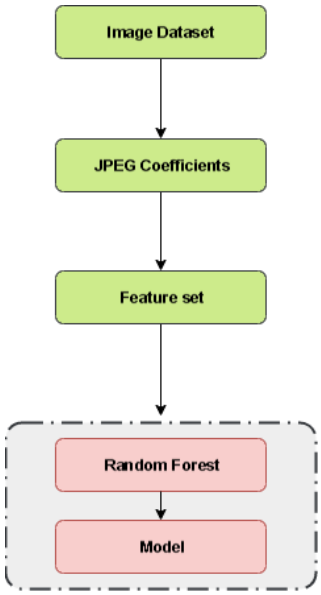

FIGURE 5. "Phase 1" of the recommendation process as a machine learning paradigm.

The model is used to find the class/type of products that are searched or clicked by the user.

For the first ML phase of a product class learning, we use the RF classifier due to its generalization capabilities and increased performance in state of the art. For learning the class of the products and for extracting image features, we use the JPEG coefficients as image features. The RF is further

integrated into the DL setup for performance analysis and performance enhancements.

Figure 5 shows the steps of JPEG features being used by the RF classifier for category learning. From Figure 5, the available image dataset is used for learning features and the type of product. The features are extracted based on the JPEG coefficients. The Coefficients are converted to a feature vector. This feature vector constitutes the feature set for the RF classifier. The RF classifier learns the distribution of the features for different categories of objects. The final model is stored for finding the corresponding category of the query image. Phase 1 in Figure 5 is the crucial step of the proposed recommendation system. The mistake in category prediction will lead to wrong retrievals in Phase 2. For a query image, of Figure 3, in Phase 1, the JPEG features are extracted. The RF classifier uses the trained model to predict the category of the query image. Once the category is retrieved, the image is processed by Phase 2.

B. "PHASE 1": DEEP FEATURES INTEGRATION

Figure 6 shows the Deep Learning-based RF (DL-RF) approach. The structure of the proposed DL-RF for feature extraction and classification is composed of many layers. The DL-RF is comprised of 5 Convolution Layers (CL). The CL is followed by 3 fully connected layers. Each layer uses a kernel for filtering dimensions. The Kernel coefficients are incrementally calculated during the training phase. The dot product is used for the input and weight vectors. Each neuron is linked to all outputs. Each layer uses the ReLU for expediting the learning process. The Softmax layer takes input from the last fully connected layer and generates a probabilistic distribution of the category of the images. Figure 6 shows the proposed architecture for enhancing the performance of the category classification of Phase 1. Features from DL are obtained from the C7 layer and RF learns the class distribution based on these DL features.

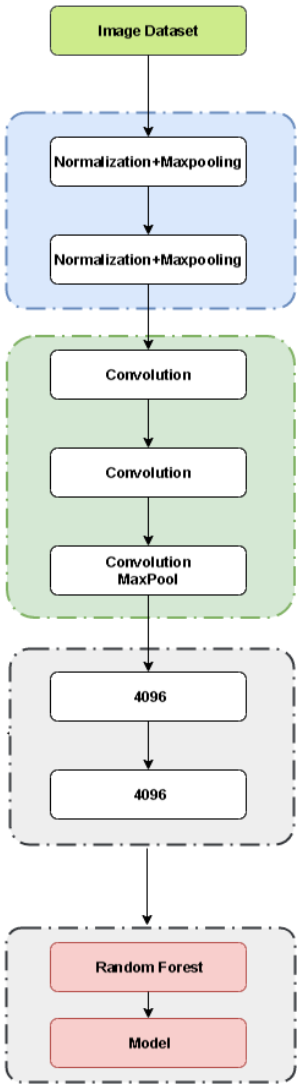

FIGURE 6. Random Forest integration into deep learning. The convolution layers (first five) are followed by two fully connected layers. The random forest learns and decides the category of the test image.

C. PHASE 2: IMAGE BASED RECOMMENDATION

Figure 7 shows the flow of Phase 2 for recommending similar products that are closely matching the query image. For a query image, the image category is retrieved by Phase 1. The query image is then searched in the corresponding category in Phase 2. It retrieves related images based on the similarity in a particular category. As shown in Figure 7, the category images are loaded with the query image. The JPEG features are extracted from all the images in a particular category. The JPEG features are extracted for the query image as well. This puts both the category images and the query image in the same vector space.

The next step is finding the similarity between the feature vector of the query image and the feature vectors of the category images. For similar products selection and vector matching, we use the Euclidean distance between all the vectors of the category images to the vector of the query image. These are represented as the similarity values of the query image with all the images in a particular category. The similarity scores are then sorted in the ascending order, and the top 20 are selected as the possible candidates for the recommendations as shown in Figure 7.

For most of the recommendation systems in the state of the art, the similarity index is the last step for recommending similar products. However, in the proposed approach, we extend the similarity by introducing image matching steps. Though vector matching retrieves similar images, however, the users are generally concerned about not only item matching but also similar color matching of items. For this reason, and for robust matching, we introduce further novel steps. After the 20 most similar vectors are retrieved, the images related to these vectors are re-loaded. Then these 20 images are analyzed by image-based similarity matching. For image-based similarity matching, we adopt the Structural Similarity Index (SSIM) of [28] to include color histograms in the matching process. This is represented as the "Struct-Hist" matching process. The SSIM is modified to include the color distribution (histogram) as follows:

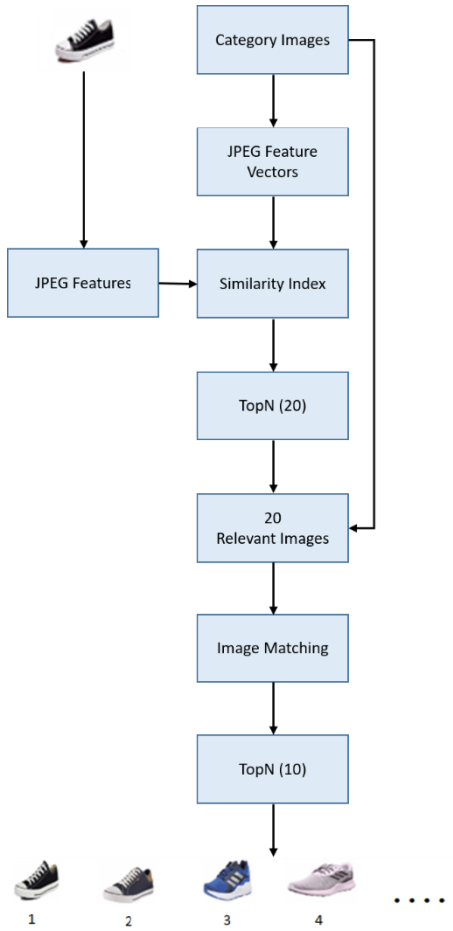

FIGURE 7. "Phase 2" of the recommendation process. The "phase 2" retrieves similar images that of the query image after the category of the product is identified by the "Phase 1". Struct-Hist is the combination of structure and color for matching.

For an image, the luminance (L), contrast (C) and structure ($Struc$) are represented as:

$$L(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \tag{3}$$

$$C(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \tag{4}$$

$$Struc(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \tag{5}$$

where,

μ_x the average of x;
μ_y the average of y;
σ_x^2 the variance of x;
σ_y^2 the variance of y and
$c_3 = {}^{c_2}/_{2}$

An image histogram is a mass probability function of the intensity values in an image. It is employed for color images, capturing the intensity distribution of the three color channels. The color histogram is defined as,

$$h_{A,B,C}(a,b,c) = N.Prob(A = a, B = b, C = c) \tag{6}$$

FIGURE 8. Example images from the dataset.

where A, B, and C are the R, G, B, and N is the total number of pixels in an image.

We combine Equations 3, 4, 5, and 6 for proposing a Structural Histogram and represent it as "Strcut-Hist". The Struct-Hist is the weighted combination of all the parameters as follows:

$$\text{Struct}-\text{Hist}(x,y) = [L(x,y)^{\alpha}.C(x,y)^{\beta}.Struc(x,y)^{\gamma}.h] \tag{7}$$

where L is the Luminance, C is Contrast, and $Struc$ is the structure parameter. The h represents the histogram and $\propto, \beta, \gamma$ are the weighting parameters that are set to 1 for giving equal importance to all the parameters. The values of the Struct-Hist are scaled between 0 and 1. A value 1 means complete matching, 0 implies no match, and values between 0 and 1 mean the similarity and or dissimilarity between two images. The Struct-Hist combines the structural properties and the color distribution in the images. It has a number of benefits. It not only retrieves similar items but weights also the items with similar colors.

In Figure 7, a total of 20 items are retrieved based on the Euclidian distance between category image vectors and the query image vector. The 10 most similarly related products out of 20 products are selected by the Struct-Hist approach. With this approach, we believe that the combination of two steps of finding the 20 similar items based on the Euclidian distance and the most relevant 10 items by Struct-Hist always produce very accurate recommendations. The Struct-Hist based recommendation is compared with the state of the art approaches in the Experimental section.

V. EXPERIMENTAL EVALUATION

In this section, we present the dataset, performance evaluation of the proposed Phase 1 and Phase 2 of the proposed recommendation system. The related parameters and factors are also analyzed.

A. DATASET AND FEATURES

We use the Amazon product image data. We select a dataset of 3.5 million products that consists of 20 categories. Figure 4 shows the distribution of the category labels of the dataset. We randomly select 100 images for each class. The dataset is available at [34]. Example images from the dataset are shown in Figure 8.

B. "PHASE 1" PERFORMANCE ANALYSIS AND EVALUATION

For Phase 1 of the proposed approach, we use the RF meta-classifier due to its generalization capabilities and increased

performance in state of the art [29]–[31]. The features learned by the RF are based on JPEG coefficients. There are multiple reasons for choosing the JPEG Coefficients as feature vectors. The JPEG coefficients are thoroughly researched for extracting important features from images for compressing and reducing their overall size. Thus the overall image quality is maintained while sufficient image size reduction is achieved. As such, these coefficients represent the critical information in the image. From the implementation point of view, the JPEG coefficients are computationally feasible and efficient due to the DCT transformation. The transformation is regarded as a set of basis functions; thus they are efficiently precomputed and stored for the re-construction of compressed images. Also, the choice of the JPEG coefficients as feature vectors is due to the pre-comparative analysis in our work. In the preliminary experiments, we compared the features of HOG, Color Layout, and auto-correlogram features. We obtained the highest performance for the JPEG features in all cases. Also, compared to the HOG, Color Layout, and auto-correlogram features, the JPEG feature calculation is much faster. Therefore, based on all these parameters, we use JPEG features.

The performance of a particular classifier depends on the nature of data and features. For performance analysis of Phase 1, we focus on the 10-folds cross-validation as a training and testing paradigm for the validation of the class prediction. The 10-folds cross-validation not only reliably learns and tests the performance of the classifier, but is also a standard approach for classifier comparison in state of the art. The 10-folds cross-validation uses 90% data for training and 10% data for testing. For 90% of training data, the model is created and tested on the 10% testing data. The testing data is un-seen data for creating the model. Therefore, this process represents the real-world scenario for new queries to the recommendation system of Phase 1. The prediction performance from the first round is stored. The process is repeated 10 times, and the average performance is calculated for all the 10 rounds. It thus removes the bias of results, and the performance can be generalized to practical applications.

For the evaluation of Phase 1, we use the Precision, Recall, and the F-measure. The precision and recall are favored over the accuracy when the classes are not balanced. Since we have unbalanced classes in the dataset, we use these parameters for evaluation. The F-measure takes both the Precision and Recall for calculation and is a reliable parameter for similar applications. The Precision is calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

where TP is True Positive and FP is *False Positive*.

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

FN is False Negative.

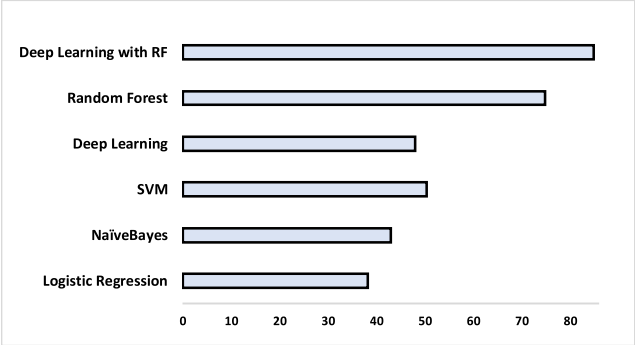

FIGURE 9. Performance evaluation of Phase 1 based on Precision. The X-axis shows the % Precision values. "Deep learning" represents the approach where features are extracted by JPEG and classified by the Deep learning. "Deep learning with RF" is the approach where the features are extracted by the Deep learning and then those features are learned by the random forest.

F-measure takes both the precision and recall into consideration as follows:

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{10}$$

The RF approach of Phase 1 is compared with the state of the art machine learning models. Following are the main comparative approaches:

Logistic Regression: is the baseline algorithm for classification and regression. It is used for multi-class classification problems [18].

Naïve Bayes: is used in many classifications and regression problems and is based on the independence property of the variables in the training data [19].

SVM: takes data to a higher dimension and uses hyperplanes to separate different classes. SVM has been used in many machine learning-based image classification tasks [20].

DL: DL algorithm is the state-of-the-art approach. DL can be used as a simple classifier, for which the non-Deep methods extract data features and DL only learns the class separation. This approach is not as robust as the one where DL is used to extract features. DL can also be used in the second approach, for which the features are learned and also classified by the DL algorithm itself. The DL in this setting is used extensively in image classification and regression problems [21]. In the comparative experiments, when we use DL, we mean the usage of DL for classifying JPEG features.

Deep Learning with RF: In this setup, DL is integrated into the RF setup. The DL learns the features using the CNN model and classifies them by the RF classifier. The RF integration into the DL setup is shown Figure 6. The 5 convolution layers are followed by two fully connected layers. After the DL features are obtained from the C7 layer, the features are then modeled by the RF classifier.

Figure 9 shows the performance evaluation of Phase 1 based on Precision. In Figure 9, the "Deep Learning" represents the approach where features are extracted by the JPEG approach and only classified by the DL. The "Deep

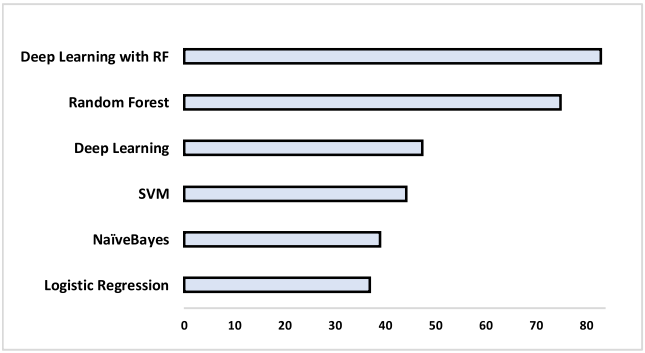

FIGURE 10. Performance evaluation of Phase 1 based on Recall. The X-axis shows the % recall values. "Deep learning" represents the approach where features are extracted by JPEG and classified by the Deep learning. "Deep learning with RF" is the approach where the and then those features are learned by the random forest.

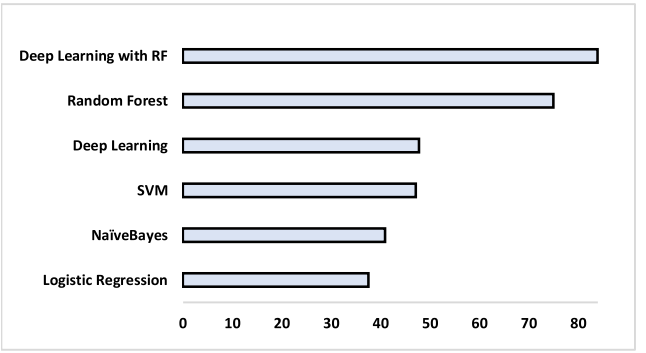

FIGURE 11. Phase 1 F-measure evaluation. The X-axis shows the % F-measure values. "Deep learning" represents the approach where features are extracted by JPEG and classified by the deep learning. "Deep learning with RF" is the approach where the features are extracted by the Deep learning and then those features are learned by the random forest.

learning with RF'' in Figure 9 depicts the proposed method in which the features are extracted by the DL and then classified and learned by the RF. From Figure 9, the Logistic Regression obtains a very low precision of 38.2%. The NaïveBayes has slightly improved precision of 43%, The SVM with precision of 50.4% is better than the Logistic Regression, Naive Bayes, and the DL (used as the Neural Network for learning class distribution only). The RF gets an increased Precision of 74.9%. The Precision of 74.9% is further enhanced to 85% by integrating the RF with the DL paradigm as shown in Figure 9.

Figure 10 shows the Recall evaluation of Phase 1. In Figure 10, the Logistic Regression gets a very low Recall of 37%, The NaïveBayes has slightly improved Recall, The SVM Recall is better than the Logistic Regression, Naive Bayes, and the DL. The RF gets an increased Recall of 75%. The Recall of 75% is further enhanced to 83% by integrating the RF with the DL approach as shown in Figure 10.

Figure 11 shows the performance evaluation of Phase 1 in terms of F-measure. In Figure 11, for the DL, features are extracted by the JPEG approach and only classified by the DL. The ''Deep learning with RF'' in Figure 11 represents the proposed approach where the features are extracted by the DL and then learned by the RF. From Figure 11, the Logistic Regression gets a very low F-measure of 37.6%. The Naïve-Bayes has slightly improved F-measure of 41%, The SVM with F-measure of 47% is better than the Logistic Regression, Naive Bayes, and the DL. The RF gets an increased F-measure of 75%. The F-measure of 75% of RF is further enhanced to 84% by integrating the RF with the DL approach, as shown in Figure 11.

C. "PHASE 2" EVALUATION

The evaluation of Phase 2 of the recommendation system is not straight forward as that of Phase 1. This is due to several reasons. First, the recommendation is a subjective process. Secondly, there is no baseline for comparison. Third, even if the data is labeled for Phase 2 of the recommendation step, it is still labeled by humans, and thus the results are subjective.

Therefore, we adopted the Autocorrelogram vector-based Euclidian distance for evaluation of Phase 2. We use the Autocorrelograms of [33] for feature vectors in Phase 2 to remove the biases that can be induced by the JPEG feature vectors used in the retrieval of Top 20 images. We could also use the JPEG vectors, but this will bias the results towards our approach because the training vectors are the JPEG vectors. Secondly, the Autocorrelograms has shown good performance for retrievability and generalization of image-based retrieval [33]. Third, the Autocorrelogram of the image captures the spatial correlation between similar intensities in the corresponding images.

For the comparative evaluation of Phase 2, we choose the K-Nearest Neighbor (KNN) and Search-based approaches. As such, we conducted experiments involving 100 query images. These images are randomly selected from the dataset. Every image is used as the query image for Phase 2 for the retrieval of ten similar images.

For evaluation, the Autocorrelograms of the ten similar images and the query images are calculated. Then each retrieved image (vector) is compared to the query image by three approaches. These comparative approaches are:

1. Subjective similarity (by 3 users)
2. Euclidian distance-based similarity
3. Cosine similarity

For subject similarity, the vectors are converted to their original images for visualization purposes and presented to users. Three students (users) perform subjective evaluation. The images are retrieved for the 100 test queries using the Struct-Hist, the KNN, and the Search-based approach. These images are then arranged with retrieved labels being removed. The labels are removed so that the user has no clue of the algorithm used for retrieval for an un-bias scoring. Each user has to give a score between 0 and 10 to the retrieval performance of the particular approach for the 10 retrieved

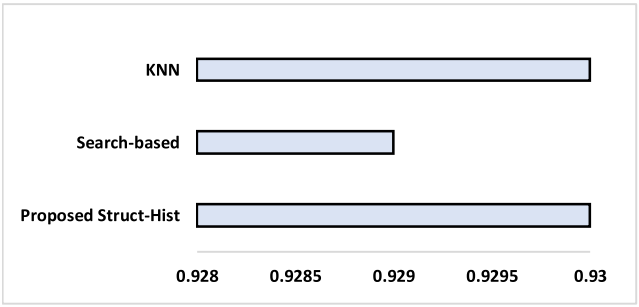

FIGURE 12. Subjective evaluation. The X-axis shows the average similarity score of test images.

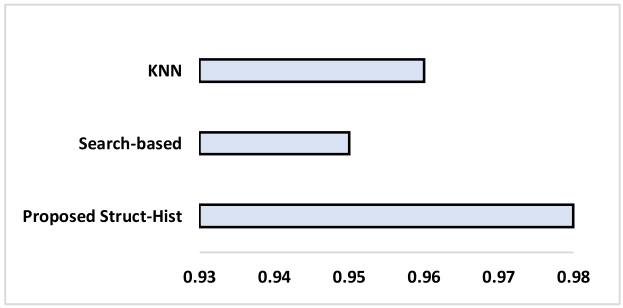

FIGURE 13. Evaluation of the recommendation phase "Phase 2" Struct-Hist step of the proposed approach by the Euclidian similarity. The X-axis shows the average similarity score of test images.

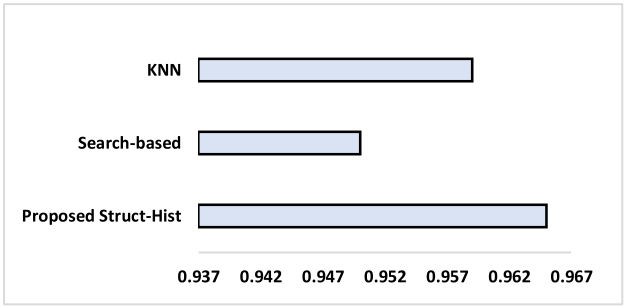

FIGURE 14. Cosine similarity. Evaluation of the recommendation phase "Phase 2" Struct-Hist step of the proposed approach by the Cosine similarity. The X-axis shows the average similarity score of test images.

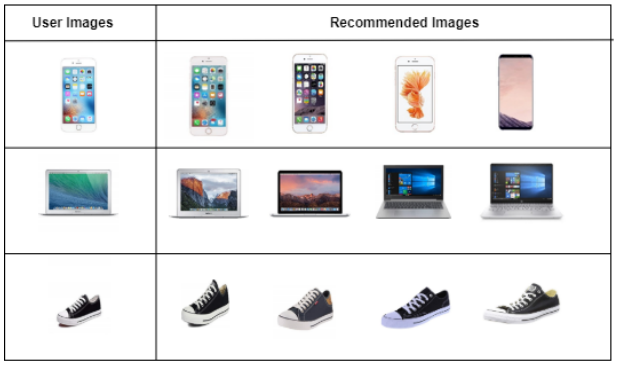

FIGURE 15. Example retrievals.

images. The scores are then averaged and scaled between 0 and 1.

Figure 12 shows the result of the comparative subjective evaluation by three users. Figure 12 shows the average subjective similarity of 0.93 for the proposed "Struct-Hist" approach. For the KNN, the average similarity is also 0.93. Search-based retrieval has an average similarity of 0.929. An interesting result is obtained in the subject evaluation. It is seen that almost all three approaches have very similar similarity scores. This can be due to the fact that human ignores some of the details of the images and focuses more on global image features. Therefore, for a particular user, it is found that the difference between an average similarity for a particular approach is not that significant. This result also shows that as long as an image is slightly related to the query image/product, the human generally considers it a complete match.

The Euclidian similarity is based on the vector (image) distance as follows:

$$Euc\_Sim = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}, \tag{11}$$

where p and q are two vectors, the Euclidian distance is then used as the similarity and scaled between 0 and 1, where 0 represents 0% similarity, and 1 means 100% similarity. All the other values between 0 and 1 indicate the corresponding percentage of similarity.

Figure 13 shows the result of the comparative evaluation using the Euclidian similarity for the three approaches. Figure 13 shows the statistics of an average similarity of 100 query images. The higher the similarity value, the better the retrieval performance. Figure 13 shows the average similarity of 0.98 for the proposed "Struct-Hist" approach. For the KNN, the average similarity is 0.96. Search-based retrieval has an average similarity of 0.95. The Euclidian similarity statistics show a good overall performance for the proposed Struct-Hist approach.

The Cosine similarity between two vectors is calculated as:

$$Cos\_sim = \frac{A.B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}, \tag{12}$$

where A_i and B_i are the components of image vectors A and B.

The Cosine similarity is scaled between 0 and 1, where 0 represents 0% similarity, and 1 means 100% similarity. All values between 0 and 1 indicate the corresponding percentage of similarity. Figure 14 shows the results of the comparative evaluation using the Cosine similarity for the three approaches. Figure 14 depicts the average similarity of 0.965 for the proposed "Struct-Hist" approach. For the KNN, the average similarity is 0.959. Search-based retrieval has an average similarity of 0.95. The Cosine similarity of Struct-Hist is higher compared to the other two approaches.

As the proposed approach consists of two steps. First retrieving 20 images, and then selecting 10 images by the proposed Struct-Hist approach. We believe that the combination

TABLE 1. Parameters and time complexity.

Parameters	Avg. Time (Seconds)
Time for calculating and learning the JPEG features by RF	**35**
Time to find the category of the query image by the JPEG and RF	**0.03**
Time for calculating and learning the Deep features by RF	**1000**
Time to find the category of the query image by the Deep features and RF	**0.3**
Time to find the top 20 related items from the dataset	**1**
Time to retrieve the 10 related items out of 20 already retrieved	**0.003**

of two steps in the proposed approach always produces accurate recommendations. The Struct-Hist approach thus outperforms the others and shows its efficacy for the image-based product recommendations. Figure 15 shows some examples of retrievals.

D. TIME COMPLEXITY

Table 1 shows the time taken by different modules. The time is represented in seconds and is the average of several runs. The experiments are conducted using the Core I7, running the Titan Nvidia GPU. Since there are many routines tested in this paper, we report the time for different modules. As in Table 1, the average time taken for calculating and learning the JPEG features by RF is 35 seconds. The time of finding the category of the query image by the JPEG and RF is 0.03 seconds. The time for calculating and learning the Deep features by RF is 1000 seconds. Time to find the category of the query image by the Deep features and RF is 0.3 seconds. The average time taken to find the top 20 related items in "Phase 2" is 1 second. Time to retrieve the 10 out of 20 associated items in "Phase 2" is 0.003 seconds.

VI. CONCLUSION

We presented an image-based product recommendation system based on two phases. In Phase 1, the proposed approach learns the class/type of the product. In Phase 2, the proposed recommendation system retrieves closely matched similar products. From the ML phase of a product's class learning, we used the RF classifier. For feature extraction from images, we used the JPEG coefficients as image features. In the evaluation of Phase 1, the proposed model generates a 75% accurate model. For performance enhancements, the RF model is further integrated into the DL setup and achieves 84% correct predictions. In Phase 2, we believe that the combination of two steps in the proposed approach of finding the 20 similar items based on the Euclidian distance and 10 most similar out of 20 by the Struct-Hist approach produces very accurate recommendations. The Struct-Hist approach thus outperforms the other methods and shows its efficacy for the image recommendation. The article contributes not only to the recommendation based systems but the algorithm presented in this article can be used for generic computer vision problems. In the future, we aim to merge non-image information with the images and use the Recurrent Neural Network (RNN) architecture for the recommendation process.

REFERENCES

[1] I. Kanellopoulos and G. G. Wilkinson, "Strategies and best practice for neural network image classification," *Int. J. Remote Sens.*, vol. 18, no. 4, pp. 711–725, 1997.

[2] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 507–517.

[3] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Image-based recommendations on styles and substitutes," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2015, pp. 43–52.

[4] V. Jagadeesh, R. Piramuthu, A. Bhardwaj, W. Di, and N. Sundaresan, "Large Scale Visual Recommendations From Street Fashion Images," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 1925–1934, 2014.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, May 2015, pp. 1–14.

[7] G. Wang, D. Hoiem, and D. Forsyth, "Learning image similarity from flickr groups using fast kernel machines," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2177–2188, Nov. 2012.

[8] G. W. Taylor, I. Spiro, C. Bregler, and R. Fergus, "Learning invariance through imitation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2729–2736.

[9] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1386–1393.

[10] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 658–666.

[11] M. Tan, S. Yuan, and Y. Su, "Content-based similar document image retrieval using fusion of CNN features," *Commun. Comput. Inf. Sci.*, vol. 819, pp. 260–270, Mar. 2018.

[12] S. Bell and K. Bala, "Learning visual similarity for product design with convolutional neural networks," *ACM Trans. Graph.*, vol. 34, no. 4, p. 98, 2015.

[13] T. Deselaers and V. Ferrari, "Visual and semantic similarity in imagenet," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1777–1784.

[14] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, Feb. 2014.

[15] J. Yang, J. Fan, D. Hubball, Y. Gao, H. Luo, W. Ribarsky, and M. Ward, "Semantic image browser: Bridging information visualization with automated intelligent image analysis," in *Proc. IEEE Symp. Vis. Analytics Sci. Technol.*, Oct./Nov. 2006, pp. 191–198.

[16] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

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

[18] L. J. Li, H. Su, E. P. Xing, and L. Fei-Fei, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1–9.

[19] S. McCann and D. G. Lowe, "Local Naive Bayes Nearest Neighbor for image classification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3650–3656.

[20] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. A. Benediktsson, "SVM- and MRF-based method for accurate classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 736–740, Oct. 2010.

[21] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.
[22] Nielsen. (2015). *Green Generation: Millennials Say Sustainability is a Shopping Priority. The Nielsen Global Survey of Corporate Social Responsibility and Sustainability*. Accessed: May 3, 2019. [Online]. Available: https://www.nielsen.com/us/en/insights/article/2015/green-generation-millennials-say-sustainability-is-a-shopping-priority/
[23] L. Liu, "Contextual topic model based image recommendation system," in *Proc. IEEE/WIC/ACM Int. Jt. Conf. Web Intell. Intell. Agent Technol. (WI-IAT)*, Dec. 2015, pp. 239–240.
[24] I. Memon, L. Chen, A. Majid, M. Lv, I. Hussain, and G. Chen, "Travel recommendation using Geo-tagged photos in social media for tourist," *Wireless Pers. Commun.*, vol. 80, no. 4, pp. 1347–1362, 2015.
[25] Y. Sun, H. Fan, M. Bakillah, and A. Zipf, "Road-based travel recommendation using geo-tagged images," *Comput. Environ. Urban Syst.*, vol. 53, pp. 110–122, Sep. 2015.
[26] L. Yu, F. Han, S. Huang, and Y. Luo, "A content-based goods image recommendation system," *Multimed. Tools Appl.*, vol. 77, no. 4, pp. 4155–4169, 2018.
[27] S. C. Guntuku, S. Roy, and L. Weisi, "Personality modeling based image recommendation," in *Proc. Int. Conf. Multimedia Modeling*, 2015, pp. 171–182.
[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[29] R. Khan, A. Hanbury, J. Stöttinger, and A. Bais, "Color based skin classification," *Pattern Recognit. Lett.*, vol. 33, pp. 157–163, Jan. 2012.
[30] R. Khan, A. Hanbury, and J. Stoettinger, "Skin detection: A random forest approach," in *Proc. Int. Conf. Image Process. (ICIP)*, Sep. 2010, pp. 4613–4616.
[31] B. Xu, Y. Ye, and L. Nie, "An improved random forest classifier for image classification," in *Proc. IEEE Int. Conf. Inf. Automat.*, Jun. 2012, pp. 795–800.
[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012.
[33] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, "Image indexing using color correlograms," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1997, pp. 762–768.
[34] *Amazon Dataset*. Accessed: Sep. 3, 2019. [Online]. Available: https://drive.google.com/drive/folders/1eaCPVw-j9ORkVHA4JY-n0QFXkHmilukm

FARHAN ULLAH received the B.Sc. degree in computer science from the University of Peshawar and the M.C.S. degree from the Sarhad University of Information Science and Technology, Peshawar. He is currently pursuing the Ph.D. degree in computer science with Shanghai University, Shanghai, China. His research interests include machine learning, recommendation systems, and computer vision.

BOFENG ZHANG received the B.S., M.S., and Ph.D. degrees from the Northwestern Polytechnical University of Technology, Xi'an, China, in 1991, 1994, and 1997, respectively. From 1997 to 1999, he worked in Zhejiang University as a Postdoctoral Researcher and was promoted to an Associate Professor. Since 1999, he has been on the faculty of the School of Computer Engineering and Science, Shanghai University. From August 2006 to August 2007, he worked as a Visiting Professor at the University of Aizu, Japan. From September 2013 to September 2014, he worked as a Visiting Professor at Purdue University Calumet, USA. He is currently the Vice Dean of the School of Computer Engineering and Science, Shanghai University. He has published three books, about 150 research articles in national and international journals and major conference proceedings. His current research interests include intelligent information processing, data science and technology, and intelligent human–computer interaction. He serves as a steering committee member, a Workshop Chair, and a program committee member of several important international conferences.

REHAN ULLAH KHAN received the B.Sc. and M.Sc. degrees in information systems from the University of Engineering and Technology, Peshawar, in 2004 and 2006, respectively, and the Ph.D. degree from the Vienna University of Technology, Austria, in 2011. He is currently an Assistant Professor with the IT Department, CoC, Qassim University, KSA. His current research interests include segmentation, machine learning, recognition, and image-based object recognition.