STYLE RETRIEVAL FROM NATURAL IMAGES

Ting-En Tseng [1], *Wei-Yi Chang* [1], *Chu-Song Chen* [2], *Yu-Chiang Frank Wang* [1]

[1] Research Center for IT Innovation, Academia Sinica, Taipei, Taiwan
[2] Institute of Information Science, Academia Sinica, Taipei, Taiwan

ABSTRACT

It has been a challenging task to identify and distinguish between images of different styles. The challenges mainly come from the extraction of high-level image semantic information, and the presence of the associated ambiguity. In this work, we propose a ranking model for style identification. Given training images of different styles, we learn a pointwise ranking model for each style based on random forests. To handle the high dimensionality of visual features and to prevent against possible ambiguity, we further introduce dimension reduction and pruning techniques for our random forests. In our experiments, we provide quantitative evaluation for style categorization in terms of mean square error (MSE) and relative ranking accuracy. Moreover, our visualization and qualitative results support the use of the proposed method for style retrieval of natural images.

Index Terms— Image Retrieval, Image Understanding, Random Forests

1. INTRODUCTION

A picture is worth a thousand words. Beyond image context, the style of an image typically plays an important role in delivering such complex ideas. In many cases, complex emotion or feeling can be conveyed with just a single and still image. As depicted in Figure 1, given images of the same scenes such as sunset and morning, different visual appearance and image composition would result in distinct image styles. While we probably do not require experts in photography to distinguish between different styles of such images, it is still a very challenging task for computers to automatically identify the image styles. Therefore, image style retrieval is the task we would like to solve in this paper. Among image styles, we consider the styles of *genre* and *mood* as suggested in [7].

Most existing works choose to address the above problem by solving a classification task. That is, given a set of images with style labels, one designs proper features or classifiers to perform style categorization. For example, Karayev *et al.* [7] considered a set of mid-level features and learned a linear classifier to separate different photographic styles. Aiming at predicting the time at which the images were taken, Palermo *et al.* [10] exploited temporal information of the extracted visual features. Xue *et al.* [13] further utilized different color-based features with Adaboost for identifying the file information of the videos (e.g., director, time, etc.).

Fig. 1. Example natural images with different styles. Note that the images in each column are of the same scene (or theme) but with distinct styles.

Nevertheless, the aforementioned works focus on style categorization, which requires and focus on a dataset with pre-collected ground truth style labels. By dividing such a dataset into training and test sets, the following task is to design feature and classifier models for recognizing the image styles. When an input image is not from the same pre-collected dataset, it is not clear whether one can directly apply the above learning models to determine its style label information. To deal with this problem, one can alternatively view style identification as a ranking problem. For example, Parikh and Grauman [11] proposed relative attribute models to observe relative ordering relationship between two images (e.g., one is more sporty than the other). They applied RankSVM as a pairwise ranking model and thus can only tell the relative information between an input pair (i.e., one is with higher or lower degree than the other in terms of the style of interest). Recently, Zhao *et al.* [14] addressed affective image retrieval by learning the optimal weight for different low and mid-level features. While image emotion can be viewed as style information, they did not utilized the degrees of each emotion when deriving their models.

In this paper, we advocate the use of a *pointwise* ranking model for style identification. We will show that, in addition to *style categorization*, the proposed model can be ap-

plied to retrieve images of particular styles with degree scores. This cannot be easily achieved by existing style classification based approaches. Our proposed ranking model is based on random forests [1], which will be extended as a regressor for ranking different image styles. We will discuss how we deal with high dimensionality of visual features and to refine the ranking scores. This is to protect the random forests against overfitting and thus for improved precision. In our experiments, we will show that our method is able to achieve improved style classification on the dataset of interest. We will further apply the proposed framework for extracting particular styles from natural images, which would be practical for applications like image retrieval, color or style transfer, and image editing.

2. OUR PROPOSED METHOD

2.1. Random Forests for Image Style Ranking

In our work, we consider the ensemble-based learning algorithm of random forests [1] for solving this task. For the completeness, we will briefly discuss the use of random forests as regressors for image ranking in Section 2.1. In Section 2.2, we will present the ideas of incorporating dimension reduction and pruning techniques in our random forests; this will be vital not only to deal with high feature dimensionality but also to refine the ranking scores.

Assume that we have N training images of the same style, and each is in terms of a feature vector $\mathbf{x} \in \mathbb{R}^d$ and a ground truth rating score y for that image style. Thus, we denote the training data as $Z_{tr} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$. This training dataset will be utilized to learn regression models with a scoring function f, so that $f(\mathbf{x}_n) \approx y_n$ can be satisfied. As a result, this training scheme can be considered as deriving a *pointwise* ranking model.

To tackle the regression problem, we need to minimize the following scoring function output, which indicates the mean squared error (MSE) loss L between predicted and ground truth values. Generally, it can be written as:

$$L(f; Z_{tr}, \mathbf{x}, y) = \frac{1}{N} \sum_{(\mathbf{x}_n, y_n) \in Z_{tr}} (f(\mathbf{x}_n) - y_n)^2. \quad (1)$$

As discussed in our experiments, we apply the AVA dataset [8] with images of different styles for training purposes. We note that, the selective image styles in our work are referred to different image categories of *mood* and *genre* determined in DPChallenge [1], not the photographic techniques (e.g., silhouettes and HDR). The ground truth scores for each image in this dataset not only reflect the image aesthetic quality, they also can be viewed as describing how the images match the associated styles. As suggested in [1, 3], among existing models for regression, random forests would be preferable in handling images with such score/degree varieties.

[1] www.dpchallenge.com

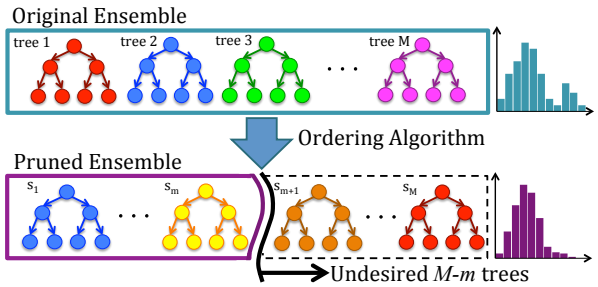

Fig. 2. Pruning of random forests. Note that m out of M outputs are preserved from the learned model to refine the prediction score. The two histograms on the right illustrate the ranking score distributions before and after pruning, respectively.

When adopting random forests as regressors for ranking, we construct M classification and regression trees (CART) [2] as base learners, each with a subset of the original N data points from Z_{tr}. This is to ensure the diversity of the individual trees in the ensemble. Let $f_i(\mathbf{x}|Z_i)$ be the response predicted by the ith tree with the corresponding sub-sampled dataset Z_i. The response of all M trees (as the output score of the random forest) will be calculated as:

$$f_{ens}^{(M)}(\mathbf{x}) = \frac{1}{M} \sum_{i=1}^{M} f_i(\mathbf{x}|Z_i). \quad (2)$$

2.2. Overfitting and Ambiguity in Image Style Ranking

2.2.1. *Protecting against overfitting via dimension reduction*

When applying the random forest as a pointwise ranking model, two additional yet practical challenges exist as we now discuss. For image processing tasks, the dimension of the feature space is typically high, which would inevitably increase the search space when constructing the trees. Moreover, for the task of image style retrieval, the image styles (and their degrees) are expected to be correlated with more than one feature attribute when traversing a tree. As a result, if one simply apply random forests as image ranking models, one might encounter potential overfitting problems.

To alleviate this problem, the technique of *dimension reduction* is introduced. In our work, we consider principal component analysis (PCA) and kernel PCA [12] on the extracted visual features as suggested by Fu *et al.* [5]. With reduced feature dimension, the split function of each node in a tree is now defined as:

$$\begin{cases} if \mathbf{F'}_i \leq threshold & \text{go to right child} \\ otherwise & \text{go to left child,} \end{cases} \quad (3)$$

where $\mathbf{F'}_i$ denotes the ith attribute in the reduced space. Later in our experiments, we will verify the effectiveness of both PCA-based techniques.

Table 1. Image styles of interest and the corresponding relative ranking accuracy obtained by our method.

Mood			
Cold	Desolation	Despair	Fear
93.51	95.51	89.63	92.73
Sacred	Sad	Silence	
91.79	93.65	95.68	
Genre			
Colors	Impressionism	Minimalism	Romance
93.62	65.15	94.49	85.71
Sepia	Vintage	Warm colors	Cool colors
94.17	88.38	92.42	91.43

2.2.2. *Protecting against ambiguity via pruning*

In contrast to overfitting, the second challenge of applying random forests as regressors is its ambiguity in determining the final output score. Recall that, despite the ability of exploiting the diversity of the observed data/features, the output is derived by averaging the scores of each individual tree. As a result, the distinctiveness between diffferent trees (and thus observed features) will be suppressed.

To disregard deviated trees in the random forest for refining the output scores, an ensemble pruning is required. While pruning techniques exist in decision tree based approaches, we need to extend such techniques for pruning our random forest (as shown in Figure 2). Recently, Hernández-Lobato *et al.* [6] proposed a greedy algorithm to identify the optimal subset from a set of regressors. Based on this idea, we perform pruning for our random forests as follows.

Given N images with feature vector $\mathbf{x}$ and a ground truth score y, the ranking score $f_i(\mathbf{x}|Z_i)$ is predicted by the ith regression tree using sub-sampled dataset Z_i. Thus, based on (1), the MSE of the final ensemble scoring function $f_{ens}^{(M)}$ with M regression trees can be rewritten as:

$$L(f_{ens}^{(M)}; Z_{tr}, \mathbf{x}, y) = \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} C_{ij}, \tag{4}$$

where $C_{ij} = \frac{1}{N} \sum_{n=1}^{N} (f_i(\mathbf{x}_n|Z_i) - y_n)(f_j(\mathbf{x}_n|Z_j) - y_n)$ denotes the covariance of the outputs between regression trees i and j. If $i = j$, C_{ii} turns into the MSE of the ith tree. For pruning purposes, we select the m trees (out of M) as the subensemble $\{s_1, s_2, ..., s_m\}$ that minimizes the loss:

$$L(f_{ens}^{(m)}; Z_{tr}, \mathbf{x}, y) = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} C_{s_i s_j}. \tag{5}$$

To select the optimal sub-ensemble for random forest pruning, we apply a sorting strategy with an ordering algorithm. Starting with an empty set, we iteratively add one regressor which reduces the MSE most. That is, for the kth iteration, the regressor to be added will satisfy:

$$s_r = \arg\min_r \frac{1}{k^2} (\sum_{i=1}^{k-1} \sum_{j=1}^{k-1} C_{s_i s_j} + 2 \sum_{i=1}^{k-1} C_{s_i r} + C_{rr}), \tag{6}$$

Table 2. Comparisons of MSE for ranking score prediction.

Linear Reg.	GBRT [4]	RF_Original Feature
1.4556	0.5703	0.5633
RF_PCA	RF_kernel PCA	Ours
0.3355	0.2950	**0.2846**

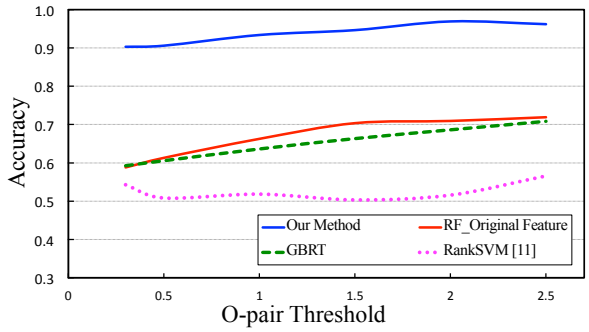

Fig. 3. Comparisons of relative ranking accuracy of different methods. Note that the horizontal axis denotes the O-pair threshold T.

where $r \in \{1, ..., N\} \setminus \{s_1, ..., s_{k-1}\}$ is the index for the remaining regression trees (not included in the subensemble, and $\{s_1, ..., s_{k-1}\}$ are those of the previously selected trees. As depicted in Figure 2, the original M trees will be sorted by its individual MSE performance, and the bottom $M - m$ ones with deviated performances will be disregarded after this pruning process. Finally, one can apply (2) with the selected m outputs for obtaining the refined output score.

3. EXPERIMENTS

We consider the AVA dataset [8], which contains 250K photographic images with a rich variety of image styles. It is worth repeating that, we consider different categories of *genre* (e.g., romance) and *mood* (e.g., sad) as the styles of interest (defined in [7]), not the photographic techniques (e.g., silhouettes and HDR), since genre and mood styles are highly related to natural images. The score of each image (as an average rating score from 1 to 10) indicates how it matches the particular image style (challenge), and also reflects the corresponding aesthetic quality. Among the images available, we choose 15 styles of interest, as listed in Table 1.

Instead of selecting sophisticated features like [7, 13] did, we extract existing standard color and appearance features as suggested [11]. To be more specific, for each image, we calculate and concatenate its 768-dimensional Lab color histogram and 320-dimensional GIST descriptor [9] as the feature representation. For the parameter selection, we fix $M = 500$ and $m = 250$ in our experiments, and the reduced feature dimension is $N - 1$ for each style category. We randomly and equally divide the images into training and test sets, and present the average results based on 15 trials.

Fig. 4. Example retrieval results. Note that the images with the highest and lowest three output scores are shown in (a) and (b), respectively.

3.1. Evaluation of Relative Ranking Accuracy

We first assess the relative ranking accuracy of our proposed method, i.e., the accuracy in determining the relative ranking relationship between two images for each style of interest. To conduct this experiment, we choose to adopt the setting of [11] in which all the image pairs (i, j) are separated into two sets: a set of ordered pairs O and the other of un-ordered pairs S. For image pairs $(i, j) \in O \implies i \succ j$, we have that image i with a stronger presence of that style than j. As for image pairs $(i, j) \in S \implies i \sim j$, we have i and j with similar relative strengths of T, which is a pre-determined threshold. In our experiment, if the difference between the predicted rating scores of images i and j is larger than T (as the ground truth difference is), then we have $i \succ j$ correctly identified for the associated style.

Table 1 shows the results of our method, in which the O-pair threshold T is set as 0.3. To show the robustness of our approach to the selection of T, we further compare with other ranking methods in Figure 3. From this figure, we see that our method was able to identify the relative ranking relationship between images, and it performed favorably against popular ranking methods. We note that, a larger O-pair threshold indicates more distinctiveness in the corresponding image style, and the resulting accuracy would improve as well. This would imply an easier retrieval task. Thus, we do not further consider the O-pair threshold beyond 3.5 in Figure 3.

3.2. Evaluation of Ranking Score Prediction

In addition to ranking accuracy, we further discuss the ability of our ranking model in determining the style degree of an image. This is measured by the mean square error (MSE) between the predicted ranking score and its ground truth value.

Several popular *pointwise* ranking models including linear regression (as a baseline) and gradient boosted regression tree (GBRT) [4] are considered for comparisons. We also have random forests using the original features, those with PCA or kernel PCA for dimension reduction, and ours without pruning as controlled experiments. Table 2 lists and compares the MSE results of different approaches. It can be seen that, the use of random forests performed favorably against linear regression and GBRT, while ours with dimension reduction and pruning further refined the predicted outputs for improved ranking performance.

3.3. Visualization of Style Retrieval

For performing style retrieval on images of different scenes, we take additional images of the AVA dataset of 6 categories: *morning, sunset, evening, landscape, skyscape* and *geology*. Take romantic sunset for example, we apply the trained random forest of style *romance*, and perform ranking score prediction on the images of *sunset* (see Figure 1). Selected visualization examples are shown in Figure 4, and more results are also available[2]. Compared to the first two parts of the experiments which focus on style categorization and score prediction, the experiments conducted in this subsection successfully support the use of our proposed method for performing style retrieval from natural images.

4. CONCLUSIONS

We presented a learning-based approach for identifying different image styles. The proposed model is based on random forests, which can be viewed as a pointwise ranking model to determine how the input images match the corresponding styles. By advancing dimension reduction and pruning techniques, we verified that our proposed model is able to alleviate potential overfitting and ambiguity when performing ranking/retrieval. Finally, quantitative and qualitative experimental results confirmed the effectiveness of our method for retrieving style information from natural images. Moreover, our method was shown to perform favorably against popular ranking approaches for style categorization.

Acknowledgement This work is supported in part by the Ministry of Science and Technology of Taiwan via MOST103-2221-E-001-021-MY2.

[2]http://styleretrieval.wix.com/styleretrieval

5. REFERENCES

[1] L. Breiman. Random forests. *Machine learning*, 2001.

[2] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC, 1984.

[3] G. Fanelli, J. Gall, and L. V. Gool. Real time head pose estimation with random regression forests. In *CVPR*, 2011.

[4] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 2001.

[5] H. Fu, Q. Zhang, and G. Qiu. Random forest for image annotation. In *ECCV*. 2012.

[6] D. Hernández-Lobato et al. Empirical analysis and evaluation of approximate techniques for pruning regression bagging ensembles. *Neurocomputing*, 2011.

[7] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller. Recognizing image style. In *BMVC*, 2014.

[8] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012.

[9] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001.

[10] F. Palermo, J. Hays, and A. A. Efros. Dating historical color images. In *ECCV*. 2012.

[11] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

[12] B. Schölkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. In *ICANN*. 1997.

[13] S. Xue, A. Agarwala, J. Dorsey, and H. Rushmeier. Learning and applying color styles from feature films. In *Computer Graphics Forum*, 2013.

[14] S. Zhao, H. Yao, Y. Yang, and Y. Zhang. Affective image retrieval via multi-graph learning. In *ACM MM*, 2014.