A DATA-DRIVEN COLOR FEATURE LEARNING SCHEME FOR IMAGE RETRIEVAL

Rahul Rama Varior, Gang Wang

Nanyang Technological University, Singapore

ABSTRACT

This paper addresses content based image retrieval based on color features. Several previous works have addressed color based image retrieval based on hand-crafted features. In this paper, a data-driven learning framework is proposed for generating color based signatures. To obtain the features, a linear transformation is learned from the pixel values based on its reconstruction error. Using this linear transformation, the original pixel values are transformed into a higher dimensional space. In the higher dimensional space, a dictionary is learned to obtain the sparse codes of the pixels. A max pooling strategy is used to obtain the dominant color features of a region and the final feature vector for an image is obtained by concatenating the pooled features. We evaluate our approach following the standard evaluation criteria for the INRIA Holidays and University of Kentucky Benchmark datasets. The approach is compared with several baselines such as histograms in RGB, HSV, YUV and Lab color spaces and several other color based features proposed for addressing this problem. Our approach shows competitive results on these datasets and outperforms all the baselines.

Index Terms— Image retrieval, color features, feature learning, data-driven framework

1. INTRODUCTION

This paper addresses the content based image retrieval problem based on color features. The objective of this problem is to retrieve images of the exact same scene or object based on a query image. A fixed size feature vector has to be computed to represent an image and further, these descriptors are used for retrieving matching images. To compute such feature descriptors based on color, we propose a learning based framework to encode each pixel in an image using a learned transformation and dictionary.

State-of-the-art algorithms for image retrieval works based on the aggregation of local descriptors, such as SIFT [1] and local color histograms. Among such holistic image descriptors, Fisher vectors [2] works best but it was recently outperformed by triangulation embedding [3]. Learned features have been used recently for image retrieval. Neural codes [4] trained on Image-Net dataset achieve state-of-the-art performance in several benchmark datasets. They follow the same architecture as mentioned in [5].

1.1. Related works on color based features

Color based features are important cues for image retrieval and yet surprisingly, color was not given much attention in retrieval tasks. A recent work [6] computes Fisher vectors based on SIFT and local color histograms and they use a metric learning framework for retrieving the images. Another major contribution towards this direction was the Bag of Colors (BOC) [7]. They compute a color codebook defined as a set of k_c colors by selecting and clustering characteristic colors from real-world images. In addition to these works, color GIST [8], color SIFT [9] and Opponent SIFT [10] were used for retrieval tasks. These color based features are mostly encoded using Bag of Words model [11]. Several other color spaces such as rg color space, Opponent color space and C color spaces were proposed for addressing photometric changes such as light intensity, shadow and shading.

The main contributions of this paper are; First, a data-driven feature learning scheme for color based features is proposed. The framework consists of a learned linear transformation to transform each pixel into a higher dimensional space and in this higher dimensional space, the sparse codes of the pixels are computed based on a dictionary. To get the dominant features over a region, a max pooling scheme is adopted. Second, we demonstrate improved search results based on color features. The approach is evaluated on the INRIA Holidays [12] and the University of Kentucky Benchmark (UKB)[13] datasets. We compare our approach with several baselines such as the color histograms in standard color spaces such as RGB, HSV, YUV and CIELab as well as several other methods mentioned in the literature for color based retrieval.

The rest of this paper is organized as follows. Section 2 explains the learning based framework for encoding the color based features. In section 3, we demonstrate the experimental evaluation of the proposed encoding scheme and compare with other competing methods. Section 4 concludes this paper.

R. Rama Varior and G. Wang are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798. E-mail: {rahul004,wanggang}@ntu.edu.sg.

2. LEARNED COLOR FEATURES

In this section, we explain the proposed data-driven feature learning and encoding scheme for color based features.

Data-driven feature learning frameworks have received a good amount of attention in the recent years. In contrast to handcrafted features such as SIFT, Local Binary Patterns and histograms, feature learning methods have shown superior performance. The implicit reason is that in feature learning methods, the patterns inherent in the data are learned or extracted automatically rather than engineering the features focusing on a particular aspect. Feature learning techniques such as auto-encoder, Restricted Boltzmann Machines (RBM) and Convolutional Neural Networks (CNN) have been used for image classification, object detection, action recognition and many other applications. Neural codes [4] have shown a detailed analysis of a CNN based feature learning framework for image retrieval and have shown that when trained with image categories belonging to the task at hand, improved search results can be obtained. However, such feature learning methods focus on extracting texture patterns and does not address the color aspects alone. A recent work [14] that addresses the illumination variation across camera views in person re-identification has proposed a learning framework to obtain illumination-invariant color features. Inspired by this work, we propose a feature learning and encoding scheme for color based features for image retrieval.

In addition to computing feature descriptors, many researchers have empirically proven that encoding techniques are also essential for achieving good performance. Even though conceptually the proposed work is close to these works, such learning based frameworks for color features have not been used in retrieval tasks. To the best of our knowledge, this is the first work that proposes a feature learning framework based on a linear transformation and dictionary to represent color based features for image retrieval.

Here after, we introduce the procedure for generating the color signature and further show that it obtains better results than several baselines and other color based features for image retrieval.

2.1. Descriptor generation

The proposed descriptor generation consists of several steps. Training data (pixels) have to be collected initially from the images and then a linear transformation is learned from these pixel values. In the linearly transformed space, a dictionary is learned and the final features, sparse codes, for an image is obtained using this dictionary. An illustration of the complete approach is shown in figure 1. Each of these steps are explained in detail below.

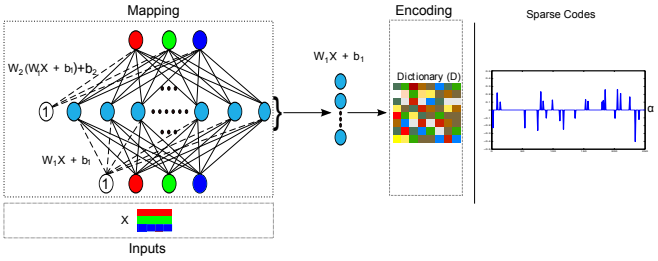

Fig. 1. Illustration of the proposed framework. Inputs are mapped into a higher dimensional space by using a learned transformation based on a single layer auto-encoder and the computed features are encoded using a dictionary. Final features are the sparse codes obtained for each pixel. Auto-encoder leads to a 60 dimensional representation and the sparse coding leads to a 250 dimensional representation for each pixel. **Best viewed in color**

2.1.1. *Training data collection*

In the proposed framework, the training data are 3-dimensional pixel values. We randomly sample 200k pixels from the entire dataset. For each image, a histogram equalization is performed for the R, G and B channels before sampling the pixels. Such randomly sampled pixels are standard normalized and thus the input data is prepared for the system.

2.1.2. *Objective formulation and optimization*

The objective of the system is to capture the stable structures and patterns inherent in the data. We extract such information from the pixels based on the reconstruction error. This concept is close to the patch based auto-encoder. The 3-dimensional pixel values are transformed to a higher dimensional space and reconstructed to the original 3-dimensional space. The objective is to minimize the reconstruction error for all the pixel values. Mathematically, it can be written as,

$$\begin{aligned} \underset{\mathbf{W_1},\mathbf{W_2},\mathbf{b_1},\mathbf{b_2}}{\text{minimize}} \quad & \frac{1}{m} \sum_{i=1}^{m} \|(W_2'(W_1'x_i + b_1) + b_2) - x_i\|_2^2 \\ + \quad & \lambda(||W_1||_F^2 + ||W_2||_F^2) \end{aligned} \tag{1}$$

where $||.||_F$ denotes the Frobenius norm. $x_i \in \mathbb{R}^{3\times 1}$ is the i^{th} pixel sampled randomly from the images and i ranges from 1 to m where m is the total number of pixels sampled. $W_1 \in \mathbb{R}^{3\times h}$, $W_2 \in \mathbb{R}^{h\times 3}$, $b_1 \in \mathbb{R}^{h\times 1}$, $b_2 \in \mathbb{R}^{3\times 1}$ are the auto-encoder parameters. h is the dimension of the projected space. All the parameters are randomly initialized and a gradient based optimization is done for the above objective function. Gradients of the objective function in equation 1 with respect to each term are computed and we use L-BFGS optimization scheme to find the local optima. Let L be the objective function in equation 1. The partial derivatives of L with

respect to each parameter are as shown below.

$$\frac{\partial L}{\partial W_1} = 2\lambda W_1 + \frac{2}{m}\sum_{i=1}^{m}((W_2'(W_1'x_i + b_1) + b_2) - x_i) \times x_i' \times W_2' \tag{2}$$

$$\frac{\partial L}{\partial W_2} = 2\lambda W_2 + \frac{2}{m}\sum_{i=1}^{m}(W_1'x_i + b_1) \times ((W_2'(W_1'x_i + b_1) + b_2) - x_i)' \tag{3}$$

$$\frac{\partial L}{\partial b_1} = \frac{2}{m}\sum_{i=1}^{m} W_2 \times ((W_2'(W_1'x_i + b_1) + b_2) - x_i) \tag{4}$$

$$\frac{\partial L}{\partial b_2} = \frac{2}{m}\sum_{i=1}^{m}((W_2'(W_1'x_i + b_1) + b_2) - x_i) \tag{5}$$

After learning the auto-encoder parameters, the inputs are transformed into the higher dimensional space by the following transformation.

$$y_i = (W_1'x_i + b_1) \tag{6}$$

After $y_i \in \mathbb{R}^{h\times 1}$ is obtained for all the training samples x_i, we compute a dictionary D to encode each y_i as sparse codes. The objective for sparse coding can be written mathematically as shown below.

$$\underset{\mathbf{D}}{\text{minimize}} \ \frac{1}{m}\sum_{i=1}^{m} ||y_i - D\alpha_i||_2^2 + \gamma||\alpha_i||_1 \tag{7}$$

where $D \in \mathbb{R}^{h\times d}$ are the basis vectors (Dictionary or Codebook) to encode each of the transformed pixel values in the higher dimensional space and d is the number of such learned dictionary elements. $\alpha_i \in \mathbb{R}^{d\times 1}$ is the sparse code generated for y_i based on the dictionary D. $||\alpha_i||_1$ enforces sparsity and γ is the penalty for the sparsity term. We use the SPAMS toolbox [15] to learn the dictionary. The proposed architecture has only one layer and the main parameters of our system are $h = 60$, $d = 250$, $\lambda = 3 \times 10^{-}3$ and $\gamma = 0.1$ as used in [14]. The time required for learning the transformation and dictionary are $3.342s$ and $27.20s$ respectively on a Core i5 3.20 GHz PC.

2.1.3. Feature computation

After learning the auto-encoder parameters and the dictionary, features are generated for each image. The images are resized to 256×256 and a histogram equalization is performed.

Table 1. Performance Comparison of different baseline color feature descriptors based on the Standard Evaluation Criteria of INRIA Holidays and UKB Dataset. Proposed Learned Color Features(LCF) outperforms all the baselines.

Method	mAP - Holidays	mAP - UKB
RGBHist	48.198	2.51
HSVHist	48.580	2.29
YUVHist	36.378	2.40
LabHist	34.097	2.14
LCF	**66.324**	**3.03**

The pixels are standard normalized and transformed to the higher dimensional space by following the equation 1. In the higher dimensional space, each of the pixels are encoded using the dictionary D following equation 7. Thus the original $256 \times 256 \times 3$ dimensional image becomes $256 \times 256 \times d$ dimensional. Further a max pooling scheme is adopted for the encoded image over 32×32, 64×64, 96×96, 112×112 and 128×128 regions. Due to the presence of multiple pooling operations at different sizes, the average time taken to compute the feature vector for each image is nearly $8.2s$. The final 23000 dimensional feature vector can be obtained by concatenating these pooled features over all the regions in the image.

2.2. Distance computation

For computing the distance between the feature vectors of each image, we follow [7] and use the L_1 distance. We also computed the L_2 distance, L_∞ and χ^2 distance, but observed that L_1 distance gave the best results. For each image, the distance between all the images are computed and the matching scores are obtained. These matching scores are used for evaluation. More details are given in the experiments section.

3. EXPERIMENTS

We evaluate our approach on INRIA Holidays dataset and the UKB dataset. The results are reported as the mean average precision (mAP). Histograms in different color spaces such as RGB (RGBHist), HSV (HSVHist), YUV (YUVHist) and Lab (LabHist) are computed on 32×32, 64×64, 96×96, 112×112 and 128×128 regions for an image and the final feature vector is the concatenation of the features computed for these regions. We compare our approach with these baselines (histograms) as well as other standard approaches and the results are as reported in the corresponding tables.

Table 2. Performance Comparison of different approaches based on color on the Standard Evaluation Criteria of INRIA Holidays Dataset. Table reports the mAP of different methods in %. D indicates the fixed size feature vector length of an image (N/A means the image is described by a set of feature descriptors). Proposed Learned Color Features(LCF) outperform the other color features based methods.

Method	Feature Descr.	D	Holidays
BOF [16]	SIFT	20000	46.9
BOF [16]	SIFT	200000	57.2
HE [16]	SIFT	N/A	74.5
colorGIST [8]	RGB Pixels	960	37.6
BOF [7]	colorSIFT	20000	57.9
HE [7]	colorSIFT	N/A	76.5
BOC [7]	Lab Pixels	1024	64.6
BOC [7]	Lab Pixels	256	63.8
LCF	RGB Pixels	23000	66.324

Table 3. Performance Comparison of different approaches based on color on the Standard Evaluation Criteria of UKB Dataset. Table reports the average retrieval score out of 4. D indicates the fixed size feature vector length of an image (N/A means the image is described by a set of feature descriptors). Proposed Learned Color Features(LCF) shows competitive performance when compared to other color based methods.

Method	Feature Descr.	D	UKB
BOF [16]	SIFT	20000	2.88
BOF [16]	SIFT	200000	2.95
HE [16]	SIFT	N/A	3.30
colorGIST [8]	RGB Pixels	960	2.06
BOF [7]	colorSIFT	20000	2.28
HE [7]	colorSIFT	N/A	2.58
BOC [7]	Lab Pixels	256	3.34
LCF	RGB Pixels	23000	3.03

3.1. INRIA Holidays dataset

The dataset consists of 1491 photographs of natural sceneries and vacation spots. There are 500 groups of images based on same scene or object. From each group, one image serves as query and the retrieval is done based on this query. The performance is reported as mAP over these 500 queries. Some of the images in the dataset are rotated by ± 90 degrees and are not in their natural orientation. We manually bring all images in the dataset to the normal orientation as done in [4]. The results are reported for both the rotated and unrotated version of this dataset. We observed an improvement of 3.186 mAP for the rotated dataset.

As shown in the table 1, the proposed encoding scheme outperforms the baselines such as the histograms in several standard color spaces. In table 2, we show the comparison of our approach with other approaches. It can be seen that our approach outperforms many standard approaches. HE [16] and HE[7] which outperforms our method for this dataset utilizes the combination of color and texture features. But it should be noted that, the proposed color based signature performs better in all the other cases and also outperforms several method that works based on a combination of color and texture features.

3.2. UKB dataset

The dataset includes 10,200 indoor photographs of 2550 objects (4 photos per object). Each image is used to query the rest of the dataset. The performance is reported as the average number of same-object images within the top-4 results, and is a number between 0 and 4. Table 1 shows the comparison of our approach against the baseline histograms and it can be seen that our approach outperforms all the baseline color features. However, from table 3, it can be seen that among other color based features, BOC [7] works better than our approach. HE [16] which outperform ours work based on a combination of color and texture. However, it should be noted that our method with color descriptors alone can outperform the results of HE [7] which is a combination of color and texture signatures.

4. CONCLUSION

We proposed a data-driven framework for learning color based features and used it to obtain the global descriptors by aggregating regional features for image retrieval. The feature descriptors are obtained by a linear transformation from raw pixel values and encoding using a dictionary in the higher dimensional space. The final feature descriptors are obtained by a max pooling scheme over regions and concatenating them. We evaluated our approach on the INRIA Holidays and UKB datasets shows that the proposed learning framework outperforms the baseline methods such as histograms in several color spaces and other methods such as the Bag of Colors and Bag of Words.

Acknowledgements: The research is supported by Ministry of Education (MOE) Tier 1 RG84/12, Ministry of Education (MOE) Tier 2 ARC28/14 and A*STAR Science and Engineering Research Council PSF1321202099.

5. REFERENCES

[1] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, "Improving the fisher kernel for large-scale image classification," in *Proceedings of the 11th European Conference on Computer Vision: Part IV*. 2010, ECCV'10, pp. 143–156, Springer-Verlag.

[3] Hervé Jégou and Andrew Zisserman, "Triangulation embedding and democratic aggregation for image search," in *CVPR - International Conference on Computer Vision and Pattern Recognition*, 2014.

[4] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky, "Neural codes for image retrieval," in *Computer Vision–ECCV 2014*, pp. 584–599. Springer, 2014.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc., 2012.

[6] A Gordoa, J.A Rodriguez-Serrano, F. Perronnin, and E. Valveny, "Leveraging category-level labels for instance-level image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),*, June 2012.

[7] Christian Wengert, Matthijs Douze, and Hervé Jégou, "Bag-of-colors for improved image search," in *MM 2011 - 19th ACM International Conference on Multimedia*. 2011, pp. 1437–1440, ACM.

[8] Aude Oliva and Antonio Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[9] Koen EA Van De Sande, Theo Gevers, and Cees GM Snoek, "Evaluating color descriptors for object and scene recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1582–1596, 2010.

[10] Joost Van De Weijer, Theo Gevers, and Andrew D Bagdanov, "Boosting color saliency in image feature detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 1, pp. 150–156, 2006.

[11] Josef Sivic and Andrew Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proceedings of Ninth IEEE International Conference on Computer Vision, 2003*.

[12] Hervé Jégou, Matthijs Douze, and Cordelia Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *European Conference on Computer Vision*. 2008, vol. I, pp. 304–317, Springer.

[13] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2006, vol. 2, pp. 2161–2168, **oral presentation**.

[14] Rahul Rama Varior, Gang Wang, and Jiwen Lu, "Learning invariant color features for person re-identification," *CoRR*, vol. abs/1410.1035, 2014.

[15] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 689–696.

[16] Hervé Jégou, Matthijs Douze, and Cordelia Schmid, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision*, vol. 87, no. 3, pp. 316–336, 2010.