Generalized Super-Vector Coding for Image Classification

M. Nakajima, Y.W. Chen & X.H. Han
Graduate School of Science and Engineering
Ritsumeikan University
Japan

Abstract — **Semantic understanding of images remains an important research challenge in machine intelligence and statistical learning. It mainly includes two steps: feature extraction and classification. This study mainly aims to explore a generalized feature extraction framework motivated by the popularly used Bag-of-feature (BOF) and super-vector coding using local descriptor, which is intuitively time –consuming for computation. In the other hand, the effortless on only exploring color and edge histogram with uniformly quantized space, which are conventional statistics, make less progress in image understanding field. Therefore, This study investigates a generalized framework based on the accessible color or edge information via adaptively modelling the explored space of a specific application, and then extracts the representation statistics (histogram of the data-driven model) and deviation statistics (the statistics of reconstruction error) for image representation. Compared to the uniformly quantized strategy such as the conventional histogram, the proposed framework can represent the image more faithful and compact, and then lead to more discriminant representation for images. With the extracted data-driven statistics, a simple linear support vector machine (SVM), which is especially efficient for large-scale database, can be effectively utilized for achieving acceptable recognition performances. Experiments on two databases: SIMPLICity and OMRON validate that our proposed strategy can achieve much better recognition performances than the conventional and the state-of-the art methods.**

Keywords-image classification; SIFT; bag-of-features; super-vector coding; support vector machine

I. INTRODUCTION

Image category recognition is important to access visual information on the level of objects and scene-like types, and it has a wide range of applications, such as intelligent image processing and content-based image indexing and retrieval (CBIR)[1,2]. Recently, the advanced progress has been made due to the explored new feature for image representation, and the progressed machine learning technique. This study mainly focuses the recognition problem of scene type images, and makes effort for the improvement of accuracy rate for scene recognition. Due to the large variation in same scene type images and the varied imaging conditions, it still remains an important research challenge in machine intelligence and statistical learning. Scene recognition mainly includes two steps: feature extraction and classification (such as Support Vector Machine and) [3] [4]. Recently, one of the most popular used image representations is Bag-of-features (BOF) with local SIFT descriptors [5], which has been proven that the acceptable recognition rates can be achieved in different application. However, it generally take a lot of computational time to extract the complex local SIFT. In the other hand, color and edge statistics such as histogram have manifested their potential performances in some applications. However, they are usually based on uniformly quantized space, and then cannot adaptively model the explored space for a specific application. Therefore, this study investigates a generalized framework based on the accessible color or edge information via adaptively modeling the explored space, and then extracts the representation statistics (histogram of the data-driven model), which would be more faithful and compact for image representation than the conventional uniformly-quantization one. In addition, integrating deviation statistics (the statistics of reconstruction error) in the data-driven model of SIFT for image representation can achieve promising performance with more compact components, which is named as Super-Vector [6]. Our proposed generalized framework can deduce not only the representation statistics but also deviation statistics of the adaptively constructed model for any explored space, and then lead to more discriminant representation for images.

II. RELATED WORKS

A. RGB Color Histogram (CH)

CH is a simple and efficient image representation method, which are very robust to any rotation, scale and translation variation in images. With all color levels, 256^3-dimensional histogram can be extracted for image representation, which will lead to high computational cost in the post-processing step. Thus, it generally reduces intensity region in each component to only a few levels such as 4~8, and obtains histogram of the reduced levels for image representation. Fig. 1 shows an example image. It is obvious from fig. 1 that Only limited quantized color levels are used to represent the original images in both 64 and 512-bin color images, which manifests the uniformly quantized color space is possibly ineffective for image representation, and then it would be more promising to adaptively learn the representative colors (codebook) for a specific application.

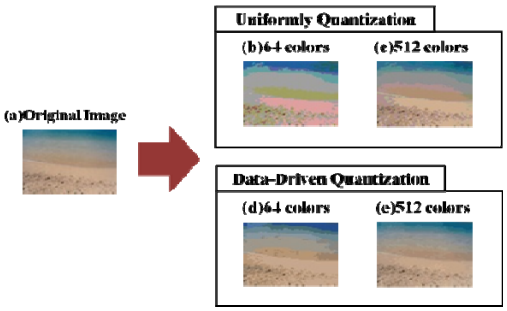

FIGURE I. THE COMPARISON OF DIFFERENT QUANTIZATION IMAGE REPRESENTATION (64 AND 512 COLORS).

B. *Edge Histogram (EH)*

EH is feature descriptors used in computer vision and image processing for the purpose of object detection. The technique counts occurrences of gradient orientation in localized portions of an image. This method is similar to that of edge orientation histograms, scale-invariant feature transform descriptors, and shape contexts, but differs in that it is computed on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved accuracy. If considering the input image as only one block, a single histogram can be extracted for image representation, and for multi-block, it just concatenates all histogram together for image representation.

C. *Bag-of-Features*

A histogram of local descriptors based on learned adaptive codebook extracted from an image. There are 4 steps in BOF: Keypoint detection, local descriptor extraction, codebook learning and histogram generation. The most popularly used local descriptor is SIFT [4], which is generally untouched in most computer vision. The SIFT descriptor computes a gradient orientation histogram within the support region (128-dimensional vector). A Gaussian window function is used to assign a weight to the magnitude of each sample point. This makes the descriptor less sensitive to small changes in the position of the support region and puts more emphasis on the gradients that are near the center of the region. It is robustness to illumination changes [8]. However, SIFT descriptor extraction is intuitively of high computational cost, which possibly limits the real application in large-scale database. With the extracted local descriptor ensemble, K-means is usually used for adaptively learning the codebook instead of uniformly-quantization, and generating the histogram for image representation as the following:

$$\arg\min_{\mathbf{V},\mathbf{U}} \sum_{m=1}^{M} \left\| \mathbf{x}_m - \mathbf{u}_m \mathbf{V} \right\|^2 \qquad (1)$$

$$\text{subject to } Card(\mathbf{u}_m) = 1, \left|\mathbf{u}_m\right| = 1, \mathbf{u}_m \geq 0, \forall m$$

where $\mathbf{V}$ is the learned codebook (prototype vector ensemble), $\mathbf{u}_m$ is the coded vector for the m^{th} descriptor $\mathbf{x}_m$. The dimension of the extracted histogram is $|C|$

III. GENERALIZED SUPER-VECTOR CODING

A. *Generalized Data-driven Model*

As mentioned in the above section, conventional CH and EH need to firstly reduce information by uniformly quantizing the raw features (color intensity or direction angle) into more compact bins due to the large intensity region, and then extracts histogram for image representation. However, for a specific recognition application, some quantized bins maybe never appear in any processed image, and at the same time the detailed variation in other quantized bins possibly include much discriminative features. Therefore, this study proposes to adaptively characterize the color ('R', 'G' and 'B') and gradient, direction angle ('θ' and 'L') information of images as for achieving more compact and effective representative vector (codebook) by K-means. With the proposed strategy, we can achieve a data driven partition of any explored raw feature space, and then extract the adaptive histogram for image representation, which is prospected to more faithful and compact than the one with the uniformly-quantized space.

Given the raw feature ensemble $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\cdots\cdots,\mathbf{x}_M]$: $\mathbf{x}_m$ can be color vector $[\mathbf{x}_m^R;\mathbf{x}_m^G;\mathbf{x}_m^B]$, gradient and direction vector $[\mathbf{x}_m^\theta;\mathbf{x}_m^L]$ and the fusion one $[\mathbf{x}_m^R;\mathbf{x}_m^G;\mathbf{x}_m^B;\mathbf{x}_m^\theta;\mathbf{x}_m^L]$, we can learn the corresponding codebook $\mathbf{V}$ with Eq. (1) for coding any raw feature sample $\mathbf{x}_t$ as the following formula:

$$C = \arg\min_{C} \left\{ \sum_{\mathbf{x}} \min_{\mathbf{v}\in C} \left\| \mathbf{x}_t - \mathbf{v} \right\|^2 \right\} \qquad (2)$$

The histogram of codebook indexes for the raw features from an image is generated as image representation. Fig.2 shows flow of our proposed method. Benefited from the effortless of obtaining the color and gradient/direction-angle vectors, the proposed generalized framework is extremely more efficient than the popularly BOF model. Fig. 1(d-e) gives an example in the adaptively learned color space (codebook size: 64, 512), which proves much more faithful representation than the conventional quantized RGB color space.

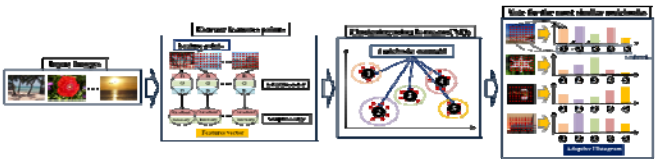

FIGURE II. FLOW OF OUR PROPOSED METHOD.

B. *Super-Vector Coding*

For a raw feature belonging to a Codebook $\mathbf{v}$, Super Vector Coding (SV) [6] adds average statistics of the differential vector between the raw feature and $\mathbf{v}$, to the Adaptive representation histogram for more well reconstruction. Given feature vector $\mathbf{x}$, where $\phi(\mathbf{x})$ is called the SV of $\mathbf{x}$, defined by

$$\phi(\mathbf{x}) = [\gamma_v(\mathbf{x}), \gamma_v(\mathbf{x})(\mathbf{x}-\mathbf{v})^{\mathrm{T}}]^{\mathrm{T}}_{v\in C}$$
$$\begin{cases} \gamma_v(\mathbf{x}) = 1 & if \quad \mathbf{v} = \mathbf{v}_*(\mathbf{x}) \\ \gamma_v(\mathbf{x}) = 0 & otherwise \end{cases} \tag{3}$$

The obtained $\phi(\mathbf{x})$ is highly sparse representation with dimension: $|C|(d+1)$, where d is the dimension of the raw feature. For example, in case of RGB color, $d=3$ (red, blue, and green value), the dimension of the final feature vector is 4|C|. In generalized adaptive model, the more codebook it has, the more well-reconstruction we can obtain. However, due to the calculation between any raw feature and all the codebook, it would be very time-consuming with large-scale codebook. Then, by integrating the (high-order) statistics of reconstruction errors, much smaller-size codebook would give more discriminated representation for image, which can greatly reduce computational cost.

IV. Experimental Results

A. *Image Database*

We validate our proposed generalized framework for image representation in two image databases. One is 'SIMPLICity Database', which consists of 10 classes with several object ones and 100 images each class (total 1000 images). The other is 'Omron Database', which consists of 8 scene classes with 200 samples (total 1600 images).

B. *Result*

For both datasets, we divide all images into 5 groups, and use 5-fold cross validation to evaluate recognition performances using different features, which means that one group is as test data and the remainder are as training data. The final recognition result is the mean of 5 groups. For SV, linear LibSVM classifier [6] is applied for achieve acceptable recognition performances, and the others are used nonlinear LibSVM classifier. Fig.3 and fig.4 show the recognition results of raw features with different codebook sizes(denoted as Cb64, Cb256, Cb512) using the conventional quantization(denoted as CH,BOF) and our proposed data-driven quantization(denoted as Ad Color, SV Color, SV Edge). Furthermore, we fuse the raw features: color and edge information, and learn the data-driven quantization to obtain statistics for image representation (Ad Color + Edge and SV Color + Edge).

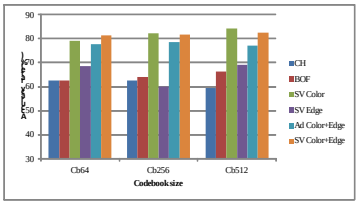

FIGURE III. Accuracy rate (SIMPLICity).

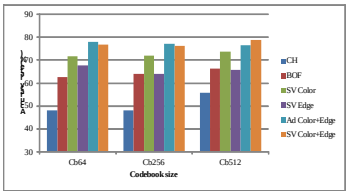

FIGURE IV. Accuracy rate (OMRON).

We can observe that the formed histogram of data-driven quantization can achieve much better recognition rates than the conventional uniformly quantization, and the combined high-order statistics (the reconstruction errors) with the histogram can give the best performance for both datasets. And by fusing two raw features, we can observe more improvement in Cb64 in SIMPLICity and Cb512 in OMRON.

In order to validate efficiency of our proposed framework, fig.5 shows the compared computational cost to the conventional BOF model which use SIFT. With the increased codebook sizes, the computational time of the conventional BOF model is exponential increased, which would prevent BOF from the real application for large-scale dataset. However, our proposed framework takes much less computational time even with large-size codebooks, which are about one-tenth of those in BOF model.

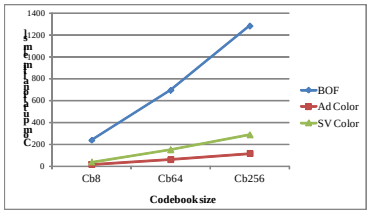

FIGURE V. Computational time.

V. Conclusion

In this paper, we proposed new data-driven model for image representation which can represent image more faithfully than the conventional color and edge histogram and intuitively be more efficient than the popular used BOF. The proposed strategy can adaptively characterize feature space for our specific application, and then achieve more discriminant features for image representation. Experiments on two databases validate that our proposed strategy can achieve much better recognition performances than the conventional and the state-of-the art methods. And we can show that computational time of our proposed method is faster than popularity used BOF. In future work, we are going to explore other coding method using raw feature, and further improve recognition performance.

References

[1] Jun Yang, Yu-Gang Jiang, Alexander Hauptmann, Chong-Wah Ngo: Evaluating Bag-of-Visual-Words Representations in Scene Classification, Carnegie Mellon University Research Showcase (2007)

[2] Yae Kikutani, Atsushi Okamoto, Xian-Hua Han, Xiang Ruan, Yen-Wei Chen: Hierarchical Classifier with Multiple Feature Weighted Fusion for Scene Recognition, Software Engineering and Data Mining (SEDM) (2010) 648--651

[3] Y.-G. Jiang, C.-W. Ngo, and J. Yang: Towards optimal bag-of-features for object categorization and semantic video retrieval, In ACM Int'l Conf, Image and Video Retrieval (2007)

[4] Yae Kikutani, Atsushi Okamoto, Xian-Hua Han, Xiang Ruan, Yen-Wei Chen: Automatic Scene Recognition by Fusion of Global and Local Feature, 2009 Joint Conference of Electrical and Electronics Engineers in Kansai (2009) G13--20 (in Japanese)

[5] H. Fujiyoshi: Gradient-Based Feature Extraction -SIFT and HOG-, Information Processing Society of Japan Research Paper CVIM160 (2007) 211--224

[6] Xi Zhou ,Kai Yu ,Tong Zhang ,Thomas S. Huang: Image Classification using Super-Vector Coding of Local Image Descriptors, ECCV 2010 Lecture Notes in Computer Science Volume 6315 (2010) 141—154.

[7] Mori, Y., Takahashi, H., and Oka, R.: Image-to-word Transformation Based on Dividing and Vector Quantizing Images with Words, International Workshop on Multimedia Intelligent Storage and Retrieval Management (1999).

[8] Jiang, Y.-G., Yang, J., Ngo, C.-W., and Hauptmann, A.G.: Representations of keypoint-based semantic concept detection: a comprehensive study, IEEE Transactions on Multimedia (2010) vol. 12 no. 1 42—53.