IMAGE RETRIEVAL USING VLAD WITH MULTIPLE FEATURES

Pin-Syuan Huang, Jing-Yi Tsai, Yu-Fang Wang, and Chun-Yi Tsai

Department of Computer Science and Information Engineering,
National Taitung University, Taiwan, R.O.C.
{u10011104,u10011127,u10011139}@ms100.nttu.edu.tw;
cytsai@nttu.edu.tw

ABSTRACT

The objective of this paper is to propose a combinatorial encoding method based on VLAD to facilitate the promotion of accuracy for large scale image retrieval. Unlike using a single feature in VLAD, the proposed method applies multiple heterogeneous types of features, such as SIFT, SURF, DAISY, and HOG, to form an integrated encoding vector for an image representation. The experimental results show that combining complementary types of features and increasing codebook size yield high precision for retrieval.

KEYWORDS

VLAD, SIFT, SURF, DAISY, HOG

1. INTRODUCTION

Image retrieval is one of the classical problems in computer vision and machine intelligence. The challenge of image retrieval is mainly aimed at trade-off between computing costs and the precision of retrieval due to the large scale data, including the large size of a single image and the large number of all images. In this paper, we propose an combinatorial encoding algorithm based on VLAD[6,9] with multiple features to achieve the goal of large scale image retrieval. VLAD cluster all features descriptors extracted from training images to find a certain number of centroids. These centroids are treated as code words or visual words, and thus form a codebook. Each image from either testing dataset or training dataset can be further encoded to a vector with fixed dimension using the trained codebook. The VLAD encoding contributes a key concept that it encodes any image by a collection of code words encoding vector with a consensus dimension. The rest of this paper is organized as follows. Section 2 reviews the previous work VLAD algorithm, section 3 elaborates the proposed method, section 4 shows experimental results, and a conclusion is given in section 5.

2. RELATED WORK

2.1. VLAD Algorithm

VLAD(vector of locally aggregated descriptors) is proposed by Jegou[6] in 2010. Similar as BOF(bag-of-features)[10], VLAD aims at representing one single image by a fixed number of feature vectors aggregated by all feature descriptors extracted from this image. Such a

David C. Wyld et al. (Eds) : ITCS, CST, JSE, SIP, ARIA, DMS - 2015
pp. 25–31, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.50103

representation is called a VLAD encoding for an image. Initially, VLAD takes all feature descriptors from all training images as inputs to cluster them and find a fixed number of centroids by K-means[7] as shown in Figure 1 and Figure 2. The collection of these centroids is referred to as the codebook. To encode an image using the codebook, the details of processing are elaborated as follows. Let N denote the total number of centroids, and c_i(i=1…N) denote the centroids.

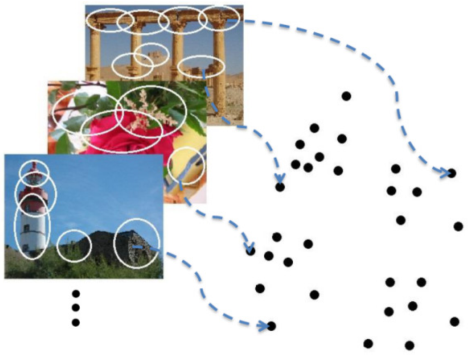

Figure 1. Features extraction from images.

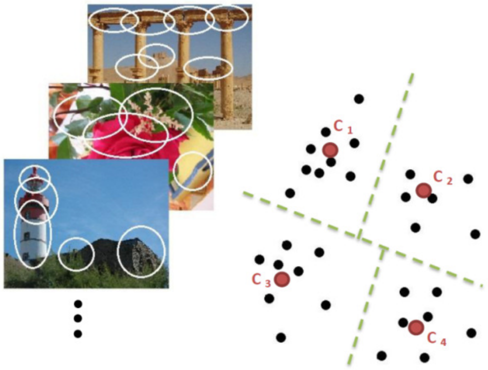

Figure 2. Clustering all features by K-means.

For encoding an image, we first extract all feature vectors from this image and denote them by x_t(t=1…T) as shown in Figure 3. Then, each feature finds the centroid closest to it, $\mathrm{NN}(x_t)$, defined in eq. (1) as shown in Figure 4.

$$\mathrm{NN}(x_t) = \arg\min ||x_t - c_i|| \qquad (1)$$

Let v_i denote the normalized vector sum of all difference between each feature vector x_t and the centroid $\mathrm{NN}(x_t)$ which it belongs to, as defined in eq. (2) and eq. (3). Then v_i(i=1…N) can be seen as an aggregation of all feature vectors contained in the input image based on the codebook as illustrated in Figure 5 and Figure 6., and such an aggregation is called a VLAD encoding for this image.

$$v_i = \sum\nolimits_{x_t:\mathrm{NN}(x_t)=c_i}(x_t - c_i) \qquad (2)$$

$$v_i := v_i/||v_i||_2 \qquad (3)$$

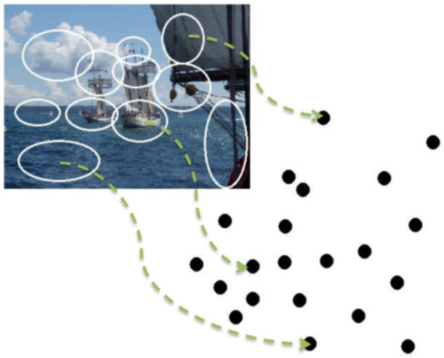

Figure 3. Features extraction for an image encoding.

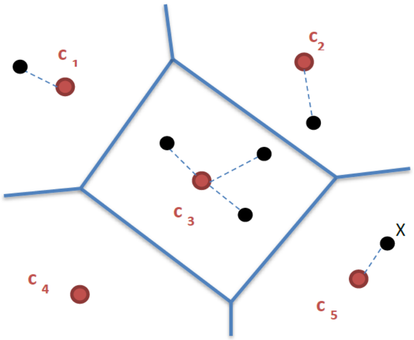

Figure 4. Each feature finds the centroid closest to it

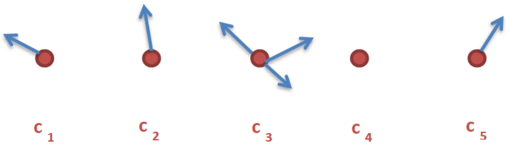

Figure 5. Distributing all features to centroids

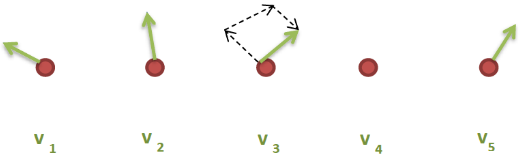

Figure 6. Vectors aggregation and normalization

3. PROPOSED METHOD

3.1. Multiple Feature Extraction

VLAD adopts SIFT as the feature descriptors for encoding. The state-of-the-art SIFT proposed by Lowe[1] is a well-known feature descriptor which is widely applied in computer vision, object recognition and machine intelligence due to its feature distinctness, robustness, and scale and rotation invariant. For achieving higher accuracy for large scale image retrieval, this paper proposes the scheme of integrating multiple types of feature into VLAD algorithm to enhance the distinctness between image objects. Thus, besides SIFT, we also adopt the well-known feature descriptors, including SURF, DAISY, and HOG. Detail of each feature descriptor is elaborated in the following.

SURF is proposed by Bay[4] which replaces the DoG in SIFT with Haar wavelets transform to generate the pyramid of scale space and approximates the determinant of Hessian blob detector by an integer evaluation for efficiency. It is feasible to be used on large scale image matching for the real-time concern. SURF descriptors can be extracted under various dimensions, for instance, SURF-64 or SURF-128 are the SURF descriptors with dimension 64 or 128, respectively.

DAISY is a local dense feature descriptor scheme proposed by Tola[2]. Similar with SIFT, it generates block-based orientation histogram but uses Gaussian convolution to aggregate these blocks for efficiency. It has been shown that DAISY is a feasible and efficient feature descriptor to be applied in wide base-line stereo matching. In this work, we uses dimension 200 to extract DAISY descriptors.

The HOG descriptor is proposed by Dalal and Triggs[8], is a well-known feature descriptor and widely applied in human or pedestrian detections due to its robustness of geometric shape and luminance. It consists of three hierarchical structures, naming cell, block, and window, respectively. The feature extraction starts by processing the image to a greyscale form, and divides the image into several cells. Each cell is partitioned into nine bins according to the orientation of gradients, and every four neighboring cells form a block. Then use a block to scan the window for each step, the length of one cell at a time. Eventually, a feature descriptor is generated by integrating feature vectors in all blocks. In this paper, we define that a cell consists of 8*8 pixels, and four cells form a block with 16*16 pixels. The block then scans a window consisting of 64*128 pixels for each step with length 8 pixels. Thus, every cell has nine orientation features, and each block contains 36 features. After window scanning is completed, we obtain (64-8)/8=7 scanning blocks in horizontal direction，and (128-8)/8=15 scanning blocks in vertical direction. Therefore, the dimension of such a HOG descriptor is 64*7*15=3780.

3.2. Codebook Training

As aforementioned, VLAD uses k-means to cluster all SIFT features extracted from all training images to find a certain number of centroids, the codebook. Similarly, we can extract other types of feature descriptors, such as SURF-64, SURF-128, DAISY, and HOG from all training images to compute the corresponding codebooks. Besides, the dimension of codebook, i.e., the number of centroids for K-means clustering is also a crucial effect upon the accuracy of recognition. In this paper, we train codebooks with 64 clusters, 128 clusters, and 256 clusters, respectively.

3.3. Combined VLAD Encoding

After finishing the codebooks generation, the encoding process for an image can be elaborated as follows. An image can be computed for its VLAD encoding vector with a specific codebook. The proposed encoding method is to combine several VLAD encoding vectors that are encoded by different codebooks into a normalized encoding vector. For instance, if four types of features, SIFT, SURF-64, DAISY and HOG are adopted and the sizes of codebook are all set to 64, an image is firstly encoded by SIFT codebook, followed by a normalization process, to produce a VLAD encoding vector of SIFT with dimension 128*64=8192. Applying similar processes, we can compute the normalized VLAD encoding vector of SURF-64, DAISY, and HOG for this image, respectively. Dimensions of SURF-64, DAISY, and HOG for a single VLAD encoding vector are 64*64=4096, 200*64=12800, and 3780*64=241920, respectively. Thus, an image can be represented by a combined normalized VLAD encoding vector with dimension 8192+4096+12800+241920=267008.

4. Experimental Results

The training and testing images adopted in this research are from INRIA Holiday dataset[5] ◦ The Holiday dataset contains 500 classes of images, and a total of 1491 images. We pick one image from each class for testing data, and the rest are for training data. Firstly, each training image is pre-processed by VLAD encoding with a specific combination of codebooks. The 500 testing image are then encoded by applying the same combination of codebook. To match the testing image with the training images, the KNN[3] algorithm is adopted to list the first 1000 ranks of encodings of training images for each encoding of testing image, and thus the mAP(mean average precision) value can be computed to reflect the precision of retrieval.

Table 1 shows the performance of encodings which combines different numbers of feature descriptors from one to four types. It is obviously that mAPs are noticeably promoted as the number of feature types and the number of centroids increases. In all combinations of any two types of feature descriptors as shown in table 2, the combination of SIFT and DAISY performs best. Our explanation is that SIFT and SURF are local descriptors, while DAISY and HOG preserves a certain portion of region description. Thus, the complementary between SIFT and DAISY can achieve more complete and detail representation of an image, and thus make each encoding more distinguishable from others. The combination of all four types of feature descriptors as shown in Table 3, SIFT, SURF(64), DAISY, and HOG, with 256 centroids yields the best result and significantly promotes the mAP to 0.72.

Table 1

Feature Descriptors	**Dimension 64-centroids**	**mAP**	**Dimension 128-centroids**	**mAP**	**Dimension 256-centroids**	**mAP**
SIFT	8192	0.528	16384	0.574	32768	0.604
SIFT+SURF(64)	12288	0.572	24576	0.597	49152	0.619
SIFT+SURF(64)+HOG	254208	0.637	508416	0.661	1016832	0.671
SIFT+SURF(64)+DAISY	25088	0.668	50176	0.679	100352	0.701
SIFT+SURF(64)+DAISY+HOG	267008	0.687	534016	0.707	1068032	**0.720**

Table 2

Feature Descriptors	Dimension 64-centroids	mAP	Dimension 128-centroids	mAP	Dimension 256-centroids	mAP
DAISY+HOG	254720	0.527	509440	0.526	1018880	0.538
SURF(128)+HOG	250112	0.566	500224	0.598	1000448	0.585
SURF(64)+HOG	246016	0.567	492032	0.585	984064	0.598
SIFT+SURF(128)	16384	0.571	32768	0.590	65536	0.607
SIFT+SURF(64)	12288	0.572	24576	0.597	49152	0.619
SURF(128)+DAISY	20992	0.594	41984	0.608	83968	0.609
SURF(64)+DAISY	16896	0.601	33792	0.607	67584	0.626
SIFT+HOG	250112	0.604	500224	0.646	1000448	0.658
SIFT+DAISY	20992	0.624	41984	0.641	83968	0.655

Table 3

Feature Descriptors	Dimension 64-centroids	mAP	Dimension 128-centroids	mAP	Dimension 256-centroids	mAP
SIFT+SURF(128)+HOG	258304	0.630	516608	0.666	1033216	0.662
SURF(64)+DAISY+HOG	258816	0.630	517632	0.642	1035264	0.654
SURF(128)+DAISY+HOG	262912	0.635	525824	0.647	1051648	0.645
SIFT+SURF(64)+HOG	254208	0.637	508416	0.661	1016832	0.671
SIFT+DAISY+HOG	262912	0.641	525824	0.667	1051648	0.675
SIFT+SURF(128)+DAISY	29184	0.661	58368	0.680	116736	0.686
SIFT+SURF(64)+DAISY	25088	0.668	50176	0.679	100352	**0.701**
SIFT+SURF(64)+DAISY+HOG	267008	0.687	534016	**0.707**	1068032	**0.720**
SIFT+SURF(128)+DAISY+HOG	271104	**0.692**	542208	**0.711**	1084416	**0.707**

5. Conclusions

In this paper, we show that a combinatorial encoding method with multiple features indeed enhance the distinctiveness among large number of image representations, and thus significantly promote the retrieval precision. On the other hand, encoding by using identical size of codebook equalizes the dimension of representation for each image, and efficiently facilitates the matching process with KNN if compared to the traditional two-stage point-to-point and geometrical matching processes for every pair of images. The benefits of distinctiveness and efficient matching make it practical and feasible to apply on large scale image retrieval. Although the dimension of encoding might increase as multiple features are applied, the computing cost caused by large dimension can be alleviated by the novel technologies of parallel processing or distributive computing.

ACKNOWLEDGEMENT

This project is fully supported by Ministry of Science and Technology in Taiwan under grant number 103-2815-C-143-008-E.

REFERENCES

[1] D. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 60(2):91–220, 2004.
[2] Engin Tola, Vincent Lepetit, Pascal Fua. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 32, Nr. 5, pp. 815 - 830, May 2010.
[3] Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 1992.
[4] Herbert Bay, Tinne Tuytelaars, Luc Van Gool. Speeded Up Robust Features. ECCV, 2006.
[5] Herve Jegou, Matthijs Douze and Cordelia Schmid. Hamming Embedding and Weak geometry consistency for large scale image search. ECCV, 2008.
[6] H.Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. CVPR, 2010.
[7] J.B. MacQueen. Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1967.
[8] N.Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. CVPR, 2005.
[9] R.Arandjelovic and A. Zisserman. All about vlad. CVPR, 2013.
[10] S.Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. CVPR, 2006.