A FAST COLOR FEATURE FOR REAL-TIME IMAGE RETRIEVAL

Chong Huang[1], Yuan Dong[1], Shusheng Cen[1], Hongliang Bai[2], Wei Liu[2], Jiwei Zhang[1], Jian Zhao[3]

[1]Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]France TelecomOrange Labs(Beijing), Beijing
[3]Patent Examination Cooperation Centers of the Patent Office, SIPO, Beijing, China
{huangchong661100, censhusheng, buptjiwei, michaeljianzhao}@gmail.com
yuan.dong@bupt.edu.cn, {hongliang.bai, wei.liu}@orange.com

Abstract: In this paper, a fast color feature is presented for real-time image retrieval. The feature is based on Dense SIFT (DSIFT) in the multi-scale RGB space. A new sum function is proposed to accelerate feature extraction instead of Gaussian weighting function. In addition, a novel randomized segment-based sampling algorithm is introduced to filter out superfluous features. In the image retrieval stage, a similarity metric is provided to measure the match between the query and reference images. After the experiments are conducted, RGB-DSIFT is more resistant to common image deformations than the original DSIFT, and more efficient than SIFT, CSIFT, GLOH feature in the processing time.

Keywords: RGB-DSIFT; Multi-scale; Color; Filter; Magnitude; Similarity measure

1 Introduction

With the development of microelectronics and computer technology, the modern computer processors get more and more powerful. Though the high operating frequency of processors gives us overall improvement for accessing visual information, global researchers are in concern with how to reduce runtime efficiently. SIFT [1] has been widely used for its robustness, but difference-of-Gaussian and other affine interest-point detectors are slow to compute. The speeded up robust feature(SURF) [2]detects keypoints by using box filters and integral images for fast computation. However, the anisotropy caused by the box filter approximation offers very low repeatability when relative direction of query and reference images varies [3]. DSIFT [4] performs well on speed without interest point detection and orientation normalization. But the number of descriptors for a medium image is very large and not all descriptors are feasible. 25572 DSIFTs are extracted when the dense sampling rate is set to 5 pixels for an896×672 pixel-wise image. If the image consists of texture-less objects, most of the descriptors are redundant and should be filtered out.

Color is a significant component for image retrieval. If it is discarded, the loss of information would increase the chances of misidentification. As Figure 1 illustrated, totally different colors seem similar after transformation from color image to gray-scale. From the point of view of information entropy theory, the information entropy loses about 66.7% in transformation from color image to gray-scale.

Figure 1 The loss of color information may affect access to visual information. Red and green windows seem similar after transformation from color image to gray-scale.

The state-of-art color descriptors vary in terms of invariant properties. The RGB histogram is statistically viewed as an approximation of an underlying continuous distribution of color values in RGB space. The opponent histogram and HSV histogram [5] are similar to the RGB histogram but different partition of color space. Color moments [6] only have shift-invariance, and moment invariants are invariant to light color intensity change and shift [5]. The Color SIFT(C-SIFT)[7] is more robust than the conventional SIFT for its photometrical-invariance. Bosch et al.[8] computes SIFT descriptors over all three channels of the HSV color model. Similarly, RGB-SIFT[9] descriptors are computed for every RGB channel independently. In [5], RGB-SIFT outperforms other features mentioned above, therefore, some characteristics and structures of the RGB-SIFT are used to reconstruct our features.

Present approaches include two kinds of features: global(gray level histograms, color histogram, etc.)and local(SIFT, SURF, etc.) features. Local features are extracted in sharp-contrast region such as corner, and global features describe an object by color signature. Though most of the latest local descriptors are preferred due to their robustness to partial appearance and their lower sensitivity to global displacements in the image, fusion between both features would give us overall improvement in image retrieval. The RGB-DSIFT and HSV-histogram are used as the basic features.

In this paper, the framework of the image retrieval system is introduced in Section 2. Section 3 describes

978-1-4673-2204-1/12/$31.00 ©2012 IEEE

the features extraction. Section 4 shows the matching and scoring. Section 5 gives the experimental results, and Section 6 gives the final conclusion of this paper and the future work.

2 System frameworks

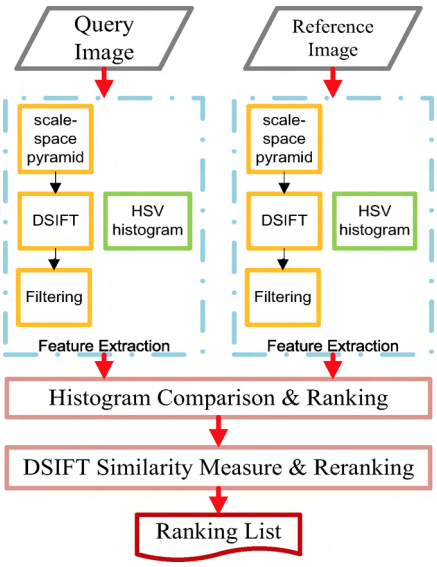

Figure 2 The framework of image retrieval system

The framework of our system is showed in Figure2. Firstly, two kinds of visual features are extracted: HSV-histogram and RGB-DSIFT descriptor. For RGB-DSIFT, the process contains three main stages: building scale-space pyramid, generating descriptor and filtering. Specially, 128D descriptor is extracted from RGB channels in the four scales respectively. All descriptors pass through the feature filtering system to reduce the redundant features. Next, the similarities among all reference against each query are computed. For each query, based on similarity of color histogram all corresponding references are ranked in descending order, and the top 150 images are selected as candidates for the next similarity measure. Then the similarities of RGB-DSIFT are calculated. In order to better evaluate the robustness of feature, the linear searching and precise matching are employed by traversing all candidates to compute the similarity of each match between descriptors from query and candidate. Finally, a score is computed for each query by the similarities of all descriptors and a ranking list is produced.

3 Feature extractions

3.1 HSV-histogram

The HSV histogram is a combination of three 1D histograms based on the Hue, Saturation, Value channels of the color space[10]. The trilinear interpolation is followed to distribute the value of each gradient sample into adjacent histogram bins. The value of each channel is quantized to 16 levels, and HSV-histogram has 48 bins.

3.2 RGB-DSIFT

The extraction of RGB-DSIFT involves three main stages: building scale-space pyramid, generating descriptor and filtering.

1) Building Scale-Space Pyramid: The spatial pyramid is proposed by Lazebnik et al, where the color space is divided into 4 octaves[11].

2) Generating Descriptor: For each channel in the RGB space, the gradient orientation and magnitude of each pixel are computed, and these samples are accumulated into orientation histogram in each sample region similar to SIFT. If the dense sampling rate is the same as the size of a region, many operations are repeated when the magnitude of each pixel over subregions is accumulated. To accelerate feature extraction, Jasper [12] proposes that two matrix multiplications are used to sum the responses within each subregion: one summarizes over the row direction, and the other does in the column direction. Given a pixel-wise responses R recording gradient orientation and magnitude of each pixel over subregions of 3×3 pixels, a matrix multiplication is employed as ARB, where A sumselements over the row direction and has the form of

$$\begin{matrix} 1 & 1 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 1 & 1 \end{matrix}$$

Matrix B sums in the column direction and is similar to transpose of A.

In terms of time complexity, the approach reduces the reiteration calculation, but uncovers such a weakness that initialization of two matrices (A and B) consumes time and memory, even if the use of specialized matrix multiplication libraries combined with sparse matrices performs better than a naive C++ implement. In addition, non-zero positions of A have a certain rule which could be used to process image.

As discussed above the following modification is provided. If the dense sampling step is set as d, the magnitude of corresponding column is summed every d rows, and the number of rows are cut by $(d-1)/d$. Similarly, the values of the corresponding row are summed every $d$columns for the previous result, and the size only occupies $1/d^2$ of its original image. Every element of matrix produced corresponds to the value of descriptor.

To avoid sudden changes in the descriptor with small changes over the sample region, the corresponding values are weighted before each summation.

3) Filtering: For the conventional DSIFT, the number of descriptors is so large that it would affect the amount of computation when finding the best matches in the next step. In particular, since multi-scale space and RGB color space are exploited, the scale of the feature would

be beyond imagination, and appropriate method is adopted to filter out some redundant descriptors.

Based on the characteristics that human vision is sensitive to the local contrast, our instinct is to take the magnitude threshold to filter the features similar to [13] because the magnitude of each descriptor describes the contrast to some extent. Before the supplementary is introduced into the system, one question is asked: whether the high-contrast areas are more informative. It is known to all that contrast ratio determines the level of details that can be seen in the image, the higher ratio providing richer color and crisper lineation. In the other words, high contrast edges help us find the color fringing. But the regions with high contrast mainly contribute to attracting attention, and do not represent the most content information. In terms of accessing the image information, regions with high contrast and low contrast are equally important.

In a different way, even if the magnitude threshold is taken as the condition for filtering values, it is a tricky problem to choose a desired threshold. On one hand, the contrast is closely related to the image content. Many details are neglected when the threshold is set as high value, and too low threshold would render the filtering system ineffective. On the other hand, information is likely to be repeated in the regions with similar contrast. Therefore, a novel filtering system is reconstructed.

In the practical system, the approach is exploited that the features are filtered randomly in each contrast level. Specially, the range of the magnitude considered is 2 to 20. The range is divided into several segments, and then the features are selected randomly in certain proportion for each segment. Selected features make up the set of RGB-DSIFT descriptors. The keypoints location of RGB-DSIFT and other features is shown in Figure 3.

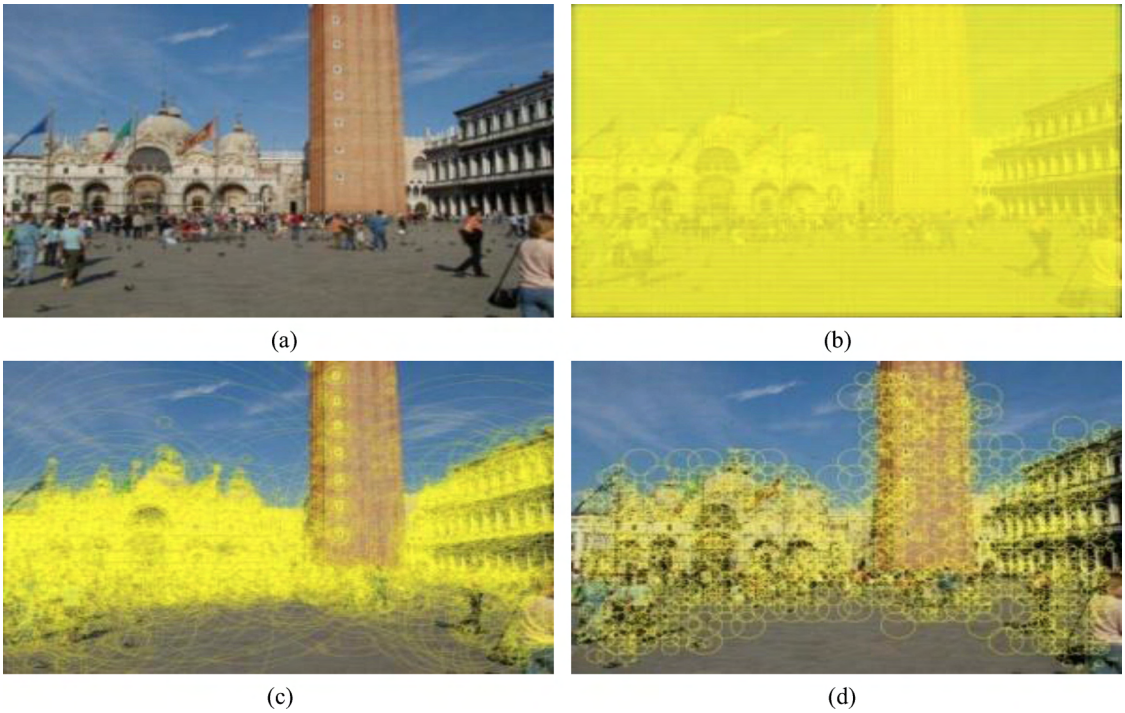

Figure 3 (a) The 896×672 original image. (b) The 25572 keypoints locations of conventional DSIFT. (c) The 5580 keypoints locations of SIFT. (d) The 2394 keypoints locations of RGB-DSIFT.

4 Matching and scoring

Considering that color histogram describes the global features and RGB-DSIFT emphasizes the local textures, the global features (HSV histogram) and local features (RGB-DSIFT) are combined to retrieve the images. For this process, 500 queries Q_i and 991 references R_j are given. Either Q_i or R_j is described by a pair {HSV histogram, RGB-DSIFTs}. Each pair from queries is sequentially compared with each pair from references.

For HSV histogram, L_1 distance is used as the similarity measure. For example, two histograms are compared as follows:

$$S_H(HSV_Q, HSV_R) = \sum_{i=1}^{3}\sum_{j=1}^{q}\left\|HSV_Q[i,j] - HSV_R[i,j]\right\| \quad (1)$$

where q is the quantization of each channel and q is set to 16.

For each descriptor from RGB-DSIFT, the distance of the closest neighbor to that of the second-closest neighbor is compared. The distance measure is set as follows:

$$d_D(DSIFT_Q, DSIFT_R) = \sum_{i=1}^{3}\sum_{j=1}^{128}\left(DSIFT_Q - DSIFT_R[i,j]\right)^2 \quad (2)$$

The matches are remained that ratio of distance (closest/second-closest) is below 0.8. In addition, if more than 3 keypoints match the same point, these matches would be abandoned. The remaining matches are illustrated as Figure 4. Query image and region are shown on the left, a matching result with the estimated corresponding region of interest is shown on the right.

For every match between a query Q_i and a reference R_j, the corresponding score is designed in the form:

$$S_D = N_{match}(1 - \frac{\sum_{i=1}^{N_{match}} d_i * ratio_i}{N_{match} * D}) \quad (3)$$

where N_{match} is the number of matches, d_i is the distance between two descriptors defined above, $ratio_i$ is the ratio of closest and second-closest distance, D is constant value and set to 24969600 (255*255*128*3) which indicates the maximum distance.

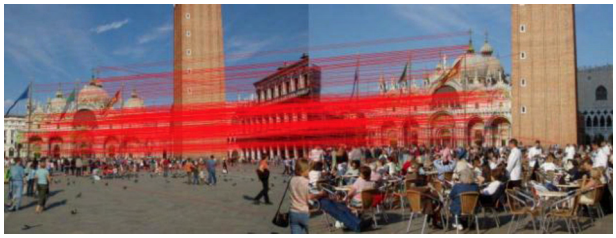

Figure 4 This example shows the positional correspondences between points on the two images. The buildings on the left image match pretty well with ones on the right image.

While performing the similarity measure, S_H is firstly computed. For each query, all corresponding references based on S_H are ranked in descending order, and the top 150 images are selected as candidates for the next similarity measure. Then all candidates are traversed to compute the S_D for every match between descriptors from query and candidate. According to the scores of each query, a ranking list is generated.

5 Experiments

5.1 Data set

The INRIA holiday dataset [14] is used as the test data. The dataset contains 500 image groups, each of which represents a distinct scene or object. The dataset includes 500 queries and 991 relevant images.

Five groups of experiments are conducted to test RGB-DSIFT. SIFT; CSIFT, GLOH, conventional DSIFT [15] and RGB-DSIFT are used as corresponding features, respectively. The same retrieval system mentioned above is adopted for all experiments.

5.2 Experimental results

Using the methods proposed above, we measure the performance of the features in terms of feature number, file size, extraction time and mAP(mean Average Precision).

As illustrated in Table 1, the size of RGB-DSIFT feature file is larger than SIFT, GLOH, and smaller than CSIFT, DSIFT. Features number is lower than that of other features. For the perspective of extraction time, as expected, runtime of RGB-DSIFT and DSIFT are the same order of magnitude, which is much faster than another three features. Due to the sensitivity to varying rotation, the mAP of RGB-DSIFT is marginally worse than CSIFT, SIFT and GLOH, but higher than DSIFT. Figure 5 gives a visual representation of retrieval performance of different features.

Table 1 Performance of different features

	CSIFT	SIFT	GLOH	DSIFT	RGB-DSIFT
Size(MB)	1221	817	901	4330	1207
Number	6.62M	6.62M	7.30M	35.4M	3.27M
Runtime	8m52s	5m29s	6m43s	3m53s	3m27s
mAP	0.666	0.661	0.697	0.580	0.643

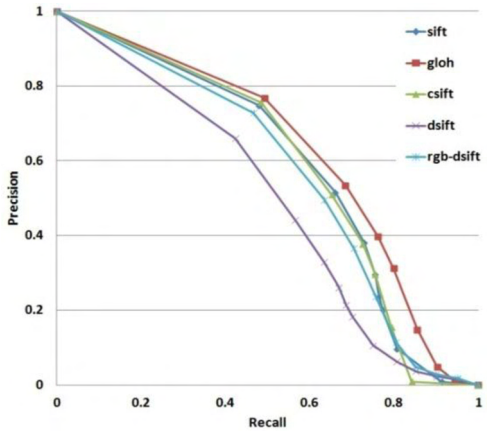

Figure 5 P-R curves comparing descriptor performance

All of the above properties explain our design decision to achieve shorter runtime by sacrificing acceptable precision, which would be efficient for time-critical application.

6 Conclusions

In this paper, a fast method of color feature extraction is presented based on DSIFT, which is practical to the real-time application. From the experimental results, we could conclude that the feature is efficient in image retrieval compared with other conventional features. In the future work, we will try to quantify the features, and pay more attention to fast retrieval algorithm to replace the linear searching.

Acknowledgements

The work is supported by Graduate Innovation Fund of SICE, BUPT, 2011. And this work is also sponsored by National Natural Science Foundation of China (90920001), and collaborative Research Project (SEV01100474) between Beijing University of Posts and Telecommunications and France Telecom R&D.

References

[1] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 60(2),91–110, 2004.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, Speeded-up robust feature, International Journal of Computer Vision, 110(3), 326-359, 2008.

[3] B.Girod, V.Chandrsekhar and D M. Chen. Mobile Visual Search. IEEE Signal Processing Magazine, 2011.

[4] L. Fei-Fei and P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. CVPR, 2005.

[5] K. E.A. van de Sande, Theo Gevers and Cees G.M. Snoek. Evaluating Color Descriptors for Object and Scene Recognition, PAMI, 32(9), 2010.

[6] F.Mindru, T. Tuytelaars L. V. Gool, and T. Moons, "Moment Invariant for Recognition under Changing Viewpoint and Illumination, CVIU,94(3) , pp.3-27,2004.

[7] A.E.Abdel-hakim and A.A.Farag, CSIFT: A SIFT Descriptor with Color Invariant Characteristics, CVPR, pp. 1978-1983, 2006.

[8] A.Bosch, A. Zisserman, and X. Munoz, Scene Classification Using a Hybrid Generative/Discriminative Approach, PAMI,50(4), 712-727, 2008.

[9] G.J. Burghouts and J.M. Geusebroek, Performance Evaluation of Local Color Invariants, CVIU, 48-62, 2009.

[10] Raphael Gonzalez, Richard E. Woods, Digital Image Processing, 2002

[11] S. Lazebnik, C. Schmid, and J.Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. CVPR, 2006.

[12] J. R.R. Uijlings, A. W.M.Smeulders and R.J.H.Scha. Real-Time Visual Concept Classification. IEEE Transactions of Multimedia, 12(7), 2010.

[13] Anna Bosch, Andrew Zisserman, Xavier Munoz. Image Classification using Random Forests and Ferns, ICCV, 2007.

[14] http://lear.inrialpes.fr/~jegou/data.php

[15] http://www.featurespace.org/