Content-based Image Retrieval Using Rotation-invariant Histograms of Oriented Gradients

Jinhui Chen[1], Toru Nakashika[1], Tetsuya Takiguchi[2], Yasuo Ariki[2]
[1]Graduate School of System Informatics, Kobe University, Kobe, 657-8501, Japan
[2]Organization of Advanced Science and Technology, Kobe University, Kobe, 657-8501, Japan
{ianchen, nakashika}@me.cs.scitec.kobe-u.ac.jp, {ariki, takigu}@kobe-u.ac.jp

ABSTRACT

Our research focuses on the question of feature descriptors for robust effective computing, proposing a novel feature representation method, namely, rotation-invariant histograms of oriented gradients (Ri-HOG) for image retrieval. Most of the existing HOG techniques are computed on a dense grid of uniformly-spaced cells and use overlapping local contrast of rectangular blocks for normalization. However, we adopt annular spatial bins type cells and apply radial gradient to attain gradient binning invariance for feature extraction. In this way, it significantly enhances HOG in regard to rotation-invariant ability and feature descripting accuracy. In experiments, the proposed method is evaluated on Corel-5k and Corel-10k datasets. The experimental results demonstrate that the proposed method is much more effective than many existing image feature descriptors for content-based image retrieval.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation

Keywords

Ri-HOG, image retrieval, content-based image category

1. INTRODUCTION

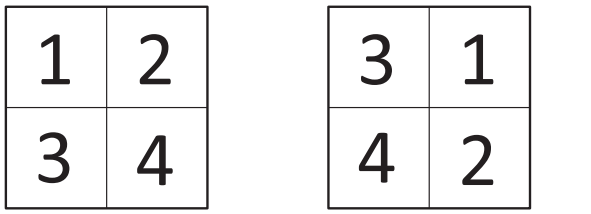
The image is one type of the most important multimedia data in human communication and it provides a rich amount of visible information for people to understand the world. With the development of Internet and digital image technologies, diverse image data appear everyday. Consequently, it has become a high demand for searching the image information that we need from the large image collections. Generally, there are main 3 types of technologies for image retrieval: the text-based, the content-based and the semantic-based. The text-based image retrieval can be traced back to 1970s. In daily life, people usually search for images mainly via Google, Yahoo, *etc.*, on text keywords, but individual perception, cognition and concept lead to many available results, which are not what they primitively needed. Therefore, in this information explosion age, textual image retrieval becomes impractical and inefficient. About the semantic-based, current image retrieval methods are usually based on low-level features (*e.g*, color, texture, spatial layout), but low-level features often fail to describe high-level concepts, *i.e.*, there is a "semantic gap" between low-level features and high-level concepts. Hence, semantic-based image retrieval is still an open problem [8]. Our research focuses on content-based image retrieval (CBIR). In this paper, we propose an effective method for integrating low-level features of images into individual perception, cognition and concept.

There are many precursors who focus on CBIR based on feature representation methods. In the MPEG-7 standard, color descriptors consist of a number of histogram descriptors, such as the dominant color descriptor, the color layout descriptor, and a scalable color descriptor [6, 5]; In the related works of CBIR, researchers also widely use texture descriptors, which provide the important information of the smoothness, coarseness and regularity of many real-world objects such as fruit, skin, clouds, trees, bricks and fabric, *etc.*, including Gabor filtering [4],local binary pattern (LBP) [7] *etc.* Generally, texture descriptors consist of the homogeneous texture descriptor, the texture browsing descriptor and the edge histogram descriptor. More recently, researchers also combine color descriptor and texture descriptor together, such as, the multi-texton histogram (MTH) [3] and the micro-structure descriptor (MSD) [2]. They use Gabor features to separately compute the color channels, in order that they can combine the color channels with classical texture descriptors to improve the feature describing ability.

In this paper, without any training procedure, clustering implementation or image segmentation, we adopt the distance metric based on a novel feature representation method for the image retrieval. The proposed feature is based on histogram descriptors. It is well known that histogram is a useful tool for image feature representation, but the robustness of many algorithms based on histogram descriptor does not reach maturity. In order to address this problem, this paper presents a new method of feature representation for CBIR, *i.e.* rotation-invariant histograms of oriented gradi-

ICMR'15, June 23–26, 2015, Shanghai, China.

http://dx.doi.org/10.1145/2671188.2749287



(a) The original block-size image (b) Image (a) rotated clockwise 90°

Figure 1: Analyzing the robustness of conventional HOG descriptors in regard to image rotation.

ents (Ri-HOG) for image retrieval. The proposed feature is reminiscent of Dalal *et al.*'s HOG [1], but the proposed feature descriptors noticeably enhance Dalal *et al.*'s version in the rotation-invariant ability and robust representation ability. We subdivide the local patch into annular spatial bins to achieve spatial binning invariance. Besides, inspired by Takacs *et al.*'s approach [9], we apply radial gradient to attain gradient binning invariance. These technical details will be discussed in Sect. 2. They are quite different from previous HOG features in the way that blocks are constructed and cells' gradient is calculated.

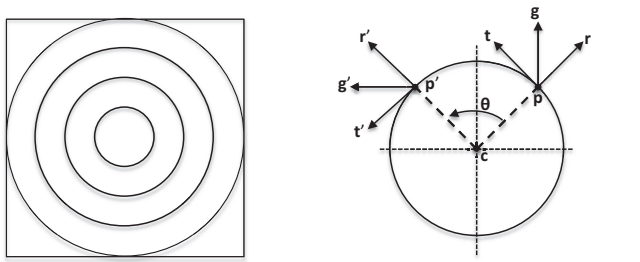
In the remainder of this paper, the Ri-HOG for image retrieval is presented in Sect. 2. Sect. 3 gives the detailed stages of the process in experimental evaluation, and conclusions are drawn in Sect. 4.

2. PROPOSED METHOD

2.1 Background and Problems

HOG are feature descriptors, which are computed on a dense grid of uniformly-spaced cells and use overlapping local contrast normalization for improved accuracy. This features set based on *cells* and *blocks* is widely used as object feature descriptors, especially the descriptors in human detection task. The describing ability of HOG features set outperforms many existing features [10]. However, the HOG feature is seldom applied to image retrieval successfully, because any feature descriptor algorithm for image retrieval must be efficient, effective and rotationally invariant. The robustness of HOG against image rotation does not reach maturity: See Fig. 1 for an example. Supposing Fig. 1(a) is a HOG block-size image, there are 4 cells in the block. Fig. 1(b) is an image of Fig. 1(a) after making a quarter turn. HOG features are extracted from the two images individually. If the histogram of oriented gradients obtained from the regions 1, 2, 3, and 4 are respectively denoted as x_1, x_2, x_3, x_4, then, the HOG features extracted from Fig. 1(a) and Fig. 1(b) are (x_1, x_2, x_3, x_4) and (x_3, x_1, x_4, x_2) respectively. This means that the rotation of image accompanies with the change of its HOG descriptors.

But why can HOG features be successfully applied to feature extraction for human detection? Because the people are usually standing, it is not necessary to worry that the orientation of task images is rotated frequently. Moreover, blocks of HOG use overlapping local contrast normalization, to some extent, which can restrain impacts brought by image rotation. Therefore, the conventional HOG just needs to be able to robustly describe the people who appear in some limited orientations and against a wide variety of background image including crowds. The high describing ability of conventional HOG makes it possible to reach this goal. However, differing from human detection situations where the people are usually standing, CBIR has to deal with the problems that the tasks in images are rotated frequently and own variable appearances. Hence, we have to substantially enhance the robustness of HOG descriptors.



(a) Annular spatial cells (b) Radial gradient transform

Figure 2: Illustration of rotation-invariant HOG descriptors.

2.2 Descriptors Extraction

Our approach: Now, the question is how to significantly improve the robustness of traditional HOG against the sharp orientation rotation of the task regions. In this paper, the answer is rotation-invariant HOG; *i.e.*, we proposed a variant of HOG owning an annular spatial cells type blocks, see Fig. 2(a). This form of blocks is reminiscent of C-HOG [1], but essentially, the approaches of C-HOG feature descriptors are the same as the R-HOG's [1]. Therefore, it cannot make HOG be rotation-invariant. Moreover, its form limits the describing ability of C-HOG. Unlike C-HOG, we use annular spatial cells to replace rectangular cells. Furthermore, these cells are computed on a dense radial gradients as feature descriptors to achieve the goal of making HOG be rotation-invariant.

How to calculate these descriptors? See Fig. 2(b), $\forall$ point p in circle c, the task is to compute the radial gradient magnitude of point p (x, y). Decompose vector g into its local coordinate system as $(g^T r, g^T t)$, by projecting g into the r and t orientations as shown in Fig. 2(b). Because the component vectors of g in r and t orientations can be obtained by $r = \frac{p-c}{\|p-c\|}$ and $t = R_{\frac{\pi}{2}} r$ quickly, and we can obtain the gradient g easily on the gradient filter. In addition, R_θ is the rotation matrix by angle θ.

Now that we have discussed the arrangement of blocks and how to obtain the feature descriptors from cells, we summarize the four steps of rotation-invariant HOG features extraction:

1. Subdivide the local patch into annular spatial cells as shown in Fig. 2(a);
2. Calculate $(g^T r, g^T t)$ of each pixel in the cell;
3. Calculate the radial gradient magnitude (M_{GR}) and its orientation (θ) on location (x, y), using the Eq. 1:

$$M_{GR}(x,y) = \sqrt{(g^T r)^2 + (g^T t)^2}, \qquad \theta(x,y) = \arctan\frac{g^T t}{g^T r}; \tag{1}$$

Table 1: Average accuracy of annuli spatial cell with different bins on Corel datasets.

Dataset	Performance	Accuracy of annular spatial cell with different bins												
		1	2	3	4	5	6	7	8	9	10	11	12	13
Corel-5K	Precision (%)	23.06	35.19	36.43	42.08	45.25	53.08	56.83	56.71	56.79	54.33	53.02	52.50	43.51
	Recall (%)	2.76	4.22	4.37	5.05	5.43	6.37	6.82	6.80	6.81	6.52	6.36	6.30	5.22
Corel-10K	Precision (%)	11.96	21.93	25.12	27.70	35.05	36.93	43.52	52.13	52.11	50.56	47.79	42.30	39.82
	Recall (%)	1.43	2.63	3.01	3.32	4.20	4.43	5.22	6.25	6.25	6.08	5.76	5.70	4.83

4. Accumulating the gradient magnitude of radial gradient for each pixel over each annular spatial cells into 8 bins.

Block normalization: We tried all of 4 normalization approaches that have been listed by Dalal *et al.* in [1]. In practice, $L_2 - Hys$, L_2 normalization followed by clipping is shown working best. The feature patch includes 10 cells and each cell contains 8 bins. The feature patch size is 100×100 pixels, which further allows different aspect ratios (the ratio of width and height). In this way, the feature can adapt various scenes and image sizes robustly. The descriptors are extracted according to the order from the inside to the outside of cells. Hence, concatenating features in 10 cells together yield a 80-dimensional feature vector.

Now, we still have the question why the Ri-HOG feature is rotation-invariant. As shown in Fig. 2 (b), assuming the local patch has been rotated by an angle ($\forall\ \theta$). Rotate point $p \rightarrow$ point p', which generates a new gradient system $R_\theta p = p';\ R_\theta r = r';\ R_\theta t = t';\ R_\theta g = g'$. We can verify the coordinates of the gradient in point p' can be expressed by $(g^T r, g^T t)$:

$$\begin{aligned}(g'^T r', g'^T t') &= ((R_\theta g)^T R_\theta r, (R_\theta g)^T R_\theta t) \\ &= (g^T R_\theta^T R_\theta r, g^T R_\theta^T R_\theta t) \\ &= (g^T r, g^T t).\end{aligned} \tag{2}$$

All rotated points in the local patch also can obtain their coordinates of the gradient from the corresponding original points, because all gradients are rotated by the same angle θ, they are one-to-one mapping. Thus, the set of gradients on any given circle or annular spatial bin centered around the patch is invariant.

3. EXPERIMENTS

In this section we will show the details of dataset, distance metric for retrieval, and evaluation results.

3.1 Experimental Dataset

The proposed method is evaluated on the Corel-5K and Corel-10K datasets [2, 3]. In the Corel-5K dataset, there are 50 categories including 5000 images, which cover diverse contents such as fireworks, bark, microscopy images, tiles, food textures, trees, waves, pills and stained glass *etc.* The Corel-10K dataset has 10,000 images covering 100 categories, such as sunsets, beaches, flowers, buildings, cars, horses, mountains, fish, food, and doors *etc.* Each category in the Corel-5K and Corel-10K datasets contains 100 JPEG images with the size of 192×128 or 128×192.

3.2 Distance Metric and Performance Metric

In this paper, the measurement of image content similarity is evaluated by distance metric. Namely, the distance between the query image and the template image in the dataset is calculated on correlation distance. The distance metric is one of the simplest approaches, therefore, it can prove the validity of descriptors most directly. Assuming the N-dimensional feature vector $T = (T_1, T_2, \cdots, T_N)$ of each template image in the dataset is extracted and stored; The feature vector of the query image is $R = (R_1, R_2, \cdots, R_N)$. Then, the correlation distance metric can be calculated as:

$$D(T, R) = \frac{\sum_{i=1}^{N} T_i R_i}{\sqrt{\sum_{i=1}^{N} (T_i)^2} \sqrt{\sum_{i=1}^{N} (R_i)^2}}. \tag{3}$$

In the experiments, N is set as 80. We use the *Precision* and *Recall* to evaluate the performance of the proposed method. These two indices are the most commonly used for evaluating image retrieval performance. Precision is the ratio of the number of retrieved similar images to the number of retrieved images; Recall is the ratio of the number of retrieved similar images to the total number of similar images. They are defined as follows:

$$\begin{aligned}P(K) &= I_K / K, \\ R(K) &= I_K / L,\end{aligned} \tag{4}$$

where L is the upper bound number of images, which are indexed from the dataset on the proposed method; K denotes the retrieval results, whose distance metrics can keep ranking in top K positions among the L similar results; I_K is the number of indexed images, whose contents are truly similar to the query image's (see the example in Fig. 3). In order to evaluate the results easily, we set $K = 12$, $L = 100$ in the same way as the setting of Liu *et al.* [2, 3]. An example of image retrieval result using the proposed method on Corel-10K is shown in Fig. 3.

Figure 3: An example of image retrieval result using the proposed method on Corel-10K (The top 12 similar template images to the query image in the dataset: the top-left image is the query image, and the images of red dashed box are the correct retrieval results).

Table 2: Average retrieval precision and recall on Corel datasets.

Dataset	Performance	Ri-HOG	HOG [1]	Gabor [4]	EHD [5]	MSD [2]	MTH [3]
Corel-5K	Precision (%)	**56.71**	41.16	36.22	39.46	55.92	49.84
	Recall (%)	**6.80**	4.91	4.35	4.74	6.71	5.98
Corel-10K	Precision (%)	**52.13**	33.29	29.15	32.31	45.62	41.44
	Recall (%)	**6.25**	3.94	3.50	3.88	5.48	4.97

3.3 Performance Evaluation

The retrieval accuracy of the proposed method not only depends on its rotation-invariant ability, but also on the number of bins in each annuli spatial cell. Generally, more spatial bins increase distinctiveness, but it will lead to the narrower annuli which decreases robustness. Therefore, there is a trade-off in performance between the number of annuli and their width, and we have to balance them. Table 1 shows the retrieval accuracies of the proposed method adopting the annulis spatial cell on different bins. We have observed that when the number of bins ≥ 7 on Corel-5K dataset and the number of bins ≥ 8 on Corel-10K dataset, the accuracy was not or seldom improved any more. But with the growth of bin number, the dimensionality of the feature also raises. Naturally, it is not beneficial to control the time complexity of feature extraction. Balancing the accuracy and time complexity for both of the database, we set the number of bins as 8. Hence, we set $N = 80$ in Subsection 3.2 and the dimensionality of the proposed method is 80.

Table 2 compares our method (Ri-HOG) with the existing image feature descriptors that were originally developed for content-based image retrieval, including Gabor features [4], the edge histogram descriptor (EHD) [5], the micro-structure descriptor (MSD) [2], and the multi-texton histogram (MTH) [3]. The Gabor features and the EHD are well-known feature representation methods for image retrieval, and the MTH and the MSD are the latest methods developed for image retrieval. These methods were conducted using their released codes. Besides, in order to evaluate the proposed method better, we also tried the conventional HOG [1] for image retrieval. Gabor features and EHD are texture descriptors, which can obtain good performance only in regular texture images. But images of real-world usually do not contain homogenous textures or regular textures, thus Gabor filter cannot represent the real-world images well. MTH and MSD combine color and texture, thus the describing ability of them is powerful. But they have ignored the local outline representation, this limits the discrimination power of them. The proposed method is a local feature, which is derived from HOG and good at describing the outline detailed information, furthermore, it has significantly enhanced the representation ability and rational robustness of HOG features. Therefore, its performance outperforms the other methods.

In addition, the average time usage of the Ri-HOG, HOG, Gabor filter, EHD, MSD and MTH are 167 ms, 183 ms, 1626 ms, 836 ms, 142 ms and 252 ms respectively (Windows 7.0 OS with Core i5 2.40 GHz CPU and 8 GB RAM).

4. CONCLUSION

In this paper, we have proposed a novel feature representation method for content-based image retrieval, *i.e.*, rotation-invariant histograms of oriented gradients (Ri-HOG). It is derived from Dalal *et al.*'s HOG yet ameliorated by the theory of polar coordinate. This histogram entirely differs from existing ones based on histograms. Furthermore it outperforms the conventional methods. This is important to those with closely related research interests. The vector dimension of the proposed feature is only 80. Therefore, it is a simple but efficient image retrieval approach. These also have been confirmed by the experiments.

About the further plan, we will try to adopt Ri-HOG as the feature representation for object recognition based on the cascade learning model.

5. REFERENCES

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 886–893, Jun. 2005.

[2] G.-H. Liu, Z.-Y. Li, L. Zhang, and Y. Xu. Image retrieval based on micro-structure descriptor. *Pattern Recognition*, 44(9):2123 – 2133, 2011.

[3] G.-H. Liu, L. Zhang, Y.-K. Hou, Z.-Y. Li, and J.-Y. Yang. Image retrieval based on multi-texton histogram. *Pattern Recognition*, 43(7):2380 – 2389, 2010.

[4] B. Manjunath and W. Ma. Texture features for browsing and retrieval of image data. *IEEE Trans. PAMI*, 18(8):837–842, Aug. 1996.

[5] B. Manjunath, J.-R. Ohm, V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Trans. Circ. Syst . Video Tech.*, 11(6):703–715, Jun. 2001.

[6] B. S. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7: multimedia content description interface*. John Wiley & Sons, 2002.

[7] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. PAMI*, 24(7):971–987, Jul. 2002.

[8] G. Papari and N. Petkov. An improved model for surround suppression by steerable filters and multilevel inhibition with application to contour detection. *Pattern Recognition*, 44(9):1999 – 2007, 2011.

[9] G. Takacs, V. Chandrasekhar, S. Tsai, D. Chen, R. Grzeszczuk, and B. Girod. Fast computation of rotation-invariant image features by an approximate radial gradient transform. *IEEE Trans. Image Proc.*, 22(8):2970–2982, Aug. 2013.

[10] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan. Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1491–1498, 2006.