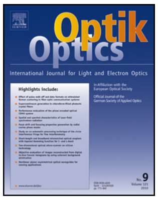
Optik 127 (2016) 141–144

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo



Similar image retrieval only using one image

Xin Meng[a], Yu An[b], Jinghan He[a], Zengqing Zhuo[c], Hao Wu[d,*], Xing Gao[b]

[a] School of Electrical Engineering, Beijing Jiaotong University, China
[b] School of Computer and Information Technology, Beijing Jiaotong University, China
[c] School of Advanced Materials, Peking University, China
[d] College of Information Science and Technology, Beijing Normal University,China

ARTICLE INFO

Article history:
Received 19 December 2014
Accepted 10 October 2015

Keywords:
Deep learning
WLS filter
Image optimization
Image decomposition

ABSTRACT

In this paper, we present one method that can retrieve the similar images only using one image. In recent years, we have used different ways to achieve image retrieval. However, if we use the unsupervised method to achieve image retrieval, the accuracy of image retrieval is reduced obviously. Even if we use the supervised method, the computing time is too long because we need to learn quite a few learning instances.

We use best feature descriptor selected, image optimization, deep learning technique to retrieve the target images that is similar to original image. On the one hand, we can see that the method makes full use of image information and select the most effective feature descriptors. On the other hand, we increase the accuracy through optimizing the target images and deep learning technique, so that it is convenient for us to extract more effective information directly. At last, we set up one big database that contains images from different categories. The images are as more complicated as possible. The experimental results show that our method not only can save the computing resource but also can keep the accuracy.

1. Introduction

In the field of computer vision, we often used the learning-based methods to achieve object recognition. However, the traditional methods are often supported by quite a few learning instances. Some previous studies [1–5] need to learn the model by using enough learning instances. Even though some of them can reduce the learning instances, the accuracy of them is still reduced obviously. The drawbacks also exist in the field of image classification [6,7], image retrieval [8], and image annotation [9,10].

In recent years, quite a few papers did some research on reducing the learning instances. Similarity information was applied to learn the model. The metric learning method is one of famous methods. Some methods [11–13] make full use of feature space and measure the semantic distance using adjusting the parameters. Then some analogous process was used to evaluate the similarity instead of metric learning [14,15]. Alternative global similarity process gets similar data using multidimensional scaling [16]. However, even some of them can achieve the high accuracy, it still needs some instances. Nearly, no paper can retrieve the image by only using one image. But in this paper, we can try to retrieve similar images without learning with SVM model.

In this paper, the feature descriptors are also important section for our model. There are numerous feature descriptors in recent years (e.g., RGB color histograms, texture histograms, SIFT [17], rgSIFT [18], PHOG [19], and GIST [20]). Not only for these methods, the other improved feature descriptors are still used [21–26]. However, we found that if we want to extract the semantic information, the GIST and SIFT are the most effective descriptors. We can also do some experiments to verify them in section 4.

At last, we set up one new database with enough images. We also did some experiments to show our method's superiority compared to some baselines.

2. Algorithm

In this section, we use best feature selected, optimized image segmentation, and deep learning to achieve our results. Firstly, we select the best feature descriptors with our method. Then, we use WLS filter to optimize the image and segment the target image. At last, we use deep learning method to extract the image information deeply.

2.1. Best feature selected

As discussed above, there are quite a few feature descriptors in the field of computer vision. However, if we use them simultaneously, they will be reused. Moreover, using them simultaneously

* Corresponding author.
E-mail address: 10112056@bjtu.edu.cn (H. Wu).

http://dx.doi.org/10.1016/j.ijleo.2015.10.041
0030-4026/© 2015 Elsevier GmbH. All rights reserved.

will add a special burden for the computer. So we need to use the most effective feature descriptors. After some experimental results, we found that some basic descriptors can determine the basic feature, such as texture, color, and brightness. However, if we want to extract the high-level semantic information, we can only use high-level feature descriptors. GIST and SIFT are classic high-level feature descriptors and they are widely used. But how to combine them or select the best one has been one challenging topic. In this paper, we use the mean shift method to set the model. Because the mean shift is widely used, so we only give one brief introduction of it. We selected some images from scenes and objects. Then we put the candidate images to the databases and used the mean shift to determine which category the candidate images belong to. If the candidate image belongs to scene category, we used the GIST descriptor to learn the model. If the candidate images belong to the object category, we used the SIFT descriptor to learn the model.

2.2. Image optimization

In this section, we first present the WLS filter, which not only can maintain sharp edges but also can smooth the regions. Then we present how to achieve multi-scale decomposition based on WLS filter. At last, we use this technique to retrieve images.

Edge preserved while region smoothed can be treated as one trade-off issue. On the one hand, the filter can make handled image as similar as original image. On the other hand, the filter can smooth the image while sharping the edges. Then the filter can be expressed by one mathematical equation as follow:

$$H(i) = \sum\nolimits_i \left(k * (a_{x,p}(i) \left(\frac{\partial u}{\partial x} \right)_i^2 + a_{y,p}(i) \left(\frac{\partial u}{\partial x} \right)_i^2 + (u_i - g_i)^2 \right), \tag{1}$$

where the pixel in the image can be defined as subscript i. The term $(u_i - g_i)^2$ was used to minimize the distance between the filtered image u and the original image g .The term $\sum_i (k * (a_{x,p}(i) \left(\frac{\partial u}{\partial x} \right)_i^2 + (a_{y,p}(i) \left(\frac{\partial u}{\partial x} \right)_i^2)$ is used for smoothing the image while sharping the significant gradients(edges). We often use the k to balance the two terms. In the process of decomposition, we can increase smoothness for each layer through changing the parameter k. The larger the k is, the smoother filtered image is:

$$a_{x,p}(i) = 1/ \left(\left| \frac{\partial l}{\partial x}(i) \right| * Z + c \right) \tag{2}$$

$$a_{y,p}(i) = 1/ \left(\left| \frac{\partial l}{\partial y}(i) \right| * z + c \right) \tag{3}$$

Next, we will solve Eq. (1), we use linear equation as follows:

$$(I + q * Lg)u = g \tag{4}$$

$$u = F_k(g) = (I + q * L_g)^{-1} g, \tag{5}$$

where q is linear, transferred by k in Eq. (1). L_g is a matrix transferred by the term $\sum_i (k * (a_{x,p}(i) \left(\frac{\partial u}{\partial x} \right)_i^2 + (a_{y,p}(i) \left(\frac{\partial u}{\partial x} \right)_i^2)$, and the linear system has already been precisely introduced in research [27]. Then the new image u will be calculated using Eq. (5).

After the introduction of WLS filter,we apply it to construct multi-scale decompositions. Compared to regular image,multi-scale decompositions will develop more edges and smooth more regions. As we know, the Laplacian pyramid is one classical pyramid that is often used for edge preservation. However, halos reduces can have negative influence on results, so that it is difficult for us to construct one multi-scale, detail preserve decomposition using our method. As discussed above,we can adjust the smooth degrees through changing the papameters k.

The decomposition is piecewise, with smooth and coarse sections, and a sequence of different images. The coarser versions are $u^1, u^2 \ldots\ldots u^k$ of the detail layers of images g and k, that are:

$$d^i = u^{i-1} - u^i \quad \text{where}\, I = 1,\, 2 \ldots k\, u^0 = g \tag{6}$$

In Eq. (6), u^k can serve as the base layer b, and u^k will become the original image g when $i = 0$. To obtain the original image g, the base image is combined with several detail layers:

$$g = b + \sum\nolimits_{i=1}^{k} d^i \tag{7}$$

Then, the most suitable new image u^i can be selected by decreasing or increasing some layers.

2.3. Deep learning

As discussed above, we said that we will select the best features descriptors from SIFT or GIST. However, after some experiments for the GIST descriptor, there is no obvious improvement with changing the parameters. But for the SIFT descriptors, different parameters have obvious differences. So we need to achieve deep learning for SIFT descriptor. In recent years, the bag-of-words representation is widely used in the field of computer vision. It can improve the object recognition results significantly, although it is not one complicated method. In this paper, we used the multi-level version of SIFT descriptor. And our features are computed in the following:

(1) We use the uniform sampling method. We select the key points from 16 × 16 pixel patches calculated over a grid spacing of 8 pixels.
(2) The feature descriptors are used by scale invariant feature transform (SIFT) descriptor in each key points.

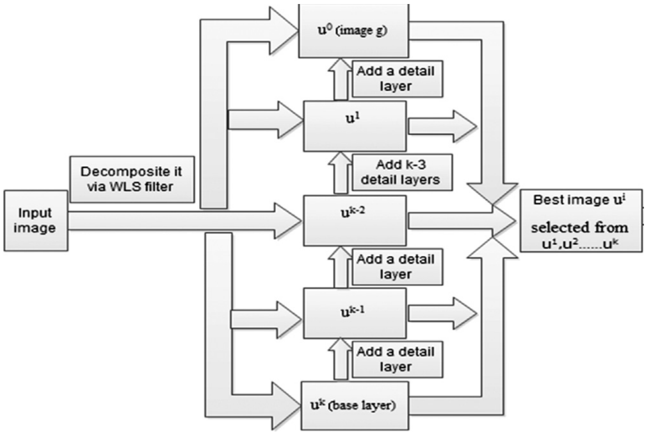


Fig. 1. The flowchart of the whole system.

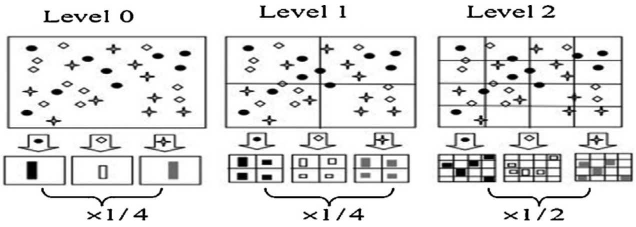


Fig. 2. One instance of constructing one three-level pyramid. The image contains three feature types, indicated by crosses, diamonds, and circles. For each level of resolution and channel,we calculate the different features which fall into each spatial bin. At last, we weigh each spatial histogram.