Adaptively Combining Local with Global Information for Natural Scenes Categorization

Shuoyan LIU\textsuperscript{(a)}, Student Member, De XU\textsuperscript{(b)}, Member, and Xu YANG\textsuperscript{(c)}, Nonmember

SUMMARY This paper proposes the Extended Bag-of-Visterms (EBOV) to represent semantic scenes. In previous methods, most representations are bag-of-visterms (BOV), where visterms referred to the quantized local texture information. Our new representation is built by introducing global texture information to extend standard bag-of-visterms. In particular, we apply the adaptive weight to fuse the local and global information together in order to provide a better visterm representation. Given these representations, scene classification can be performed by pLSA (probabilistic Latent Semantic Analysis) model. The experiment results show that the appropriate use of global information improves the performance of scene classification, as compared with BOV representation that only takes the local information into account.

key words: scene classification, pLSA, bag-of-visterms

1. Introduction

Scene classification is a very active and still challenging problem in computer vision. Generally, we divide this problem into two critical components: computation of the feature vector representing an image, and the classifier. In this paper, we mainly focus on the scene representation. Early work used low-level features. More recent approaches made use of an intermediate representation. One of the most popular representations is bags-of-visterm (BOV) \cite{1}, where visterms referred to the quantized local texture information. In the scenario of scene classification, though the general approach performed surprisingly well when only local invariant features (sparse descriptors) were used, their weakness was also derived from being local. For some cases (e.g., countryside scene), the gist of an image is not captured only through the interest points/regions. Though such an image description has been extended in many ways, it typically considers only local descriptors to create the final image representation, ignoring the other disinterest but gist regions of an image. \cite{2} combined color with texture information computed over interest points/regions in order to overcome the shortage of these descriptors. More recently, Pedro \cite{3} used multi-level local interest descriptors quantization to create a more complete representation of an image. Anna \cite{1} and Fei-fei Li \cite{4} applied dense descriptors to describe images for conquering this fault and the representation was proven to be useful for scene classification. We hope to solve the problem from another aspect by introducing global information. Using global information for image descriptors has a long history in computer vision. Aditya \cite{5} showed that global edge features are useful for classifying city and landscape. However, not much attention has been paid to combining local and global information for scene classification. One effort in this direction is proposed by \cite{7}. This paper proposed some psychophysical experiments that aim to make a further step in this direction by investigating the processing of global and local information in scene categorization. The motive for our work inspired by the result of these experiments that humans integrate both types of information together for intact scene categorization, and it has been proven by our classification experiment result (Table 1). In \cite{7} the local information was obtained by dividing an image into local regions and classifying them into local concept classes. However, our approach uses the histogram of quantized local SIFT (Scale Invariant Feature Transform) descriptors over the interest points automatically detected. For global information, distinct from \cite{7}, in which it is based on Oliva\cite{6}, we extract an edge direction histogram from the whole image. Finally, they combined the outcomes of the global and the local classifiers using a naive Bayes classifier. The limitation of \cite{7} is that it ignores the relationship between the global and local information in scene categorization. The novelty of our work, by contrast, is the adaptively weighted concatenation of the global features and local visterms. With this enlarged visual vocabulary, scene classification task is done by introducing a classification algorithm based on a combination of unsupervised probabilistic Latent Semantic Analysis (pLSA) \cite{8} followed by a discriminative classifier K nearest neighbors scheme (KNN).

2. Extended Bags-of-Visterm Scene Representation

The proposed approach may be viewed as an extended bags-of-visterms (EBOV). In fact, it is a great challenge to find a way to combine the global and local information, and expressed different meanings. But since they are measured by the same meter (i.e. histogram), this combination is possible. Figure 1 shows the basic architecture of our proposal, and each stage is described at the following sections. The input is a digital image of natural scene. Two sets of descriptors are obtained from the image: the first set corresponds to the bag-of-visterms representation of visual scenes extracted through the interest points; the second set is global informa-
tion acquired by the entire image. This raise the question of how to extend the BOV features used the global information. This question is solved by concatenating them by the optimal weight $w_d$. The optimal $w_d$ is adaptively varying according to different images, showing an emphasis on either the SIFT feature or global information. The main step for EBOV is to determine the weight value for each image based on the content. This process starts by partitioning an image into $K \times K$ regions and computing a histogram, where each bin in the histogram represents the number of interest points within a certain region. The optimal weight $w_d$ is obtained by quantized the entropy of the histogram. The next step of EBOV exploits the global information to extend bag-of-visters for all images using the suitable weights $w_d$. The result of classification is can be acquired with the help of pLSA algorithm by introducing the new extension representation. We will elaborate the framework in the following.

2.1 Bag-of-Visters Representation from Local Descriptors

The construction of the bag-of-visters vector for an image involves the following four steps [4]: (i) Automatically detect interest points/regions in an image, (ii) compute local descriptors as visual vocabulary over those interest regions/points, (iii) quantize the descriptors into visters form the visual vocabulary, (iv) count the number of occurrences of each specific word in the vocabulary in order to build the image bag-of-visters (histogram of visters). The BOV representation of the image is constructed from the local descriptors according to

$$h(d) = (h_i(d))_{i=1..N}$$

$$h_i(d) = n(d, v_i)$$

where $n(d, v_i)$ denotes the number of occurrences of vister $v_i$ in the image $d$. In this work, the local descriptors are SIFT (Scale Invariant Feature Transform), representing local texture/structure information. SIFT features are local histograms of edge directions computed over the interest points that selected by the difference of Gaussians (DOG) point detector. [9] shows that the use of 8 orientation directions and a $4 \times 4$ grid gives a good compromise between descriptor size and accuracy of representation. The final feature size is thus 128.

2.2 Global Information with Edge Direction Histograms

The edge direction is an important feature for natural scenes. For instance, the city scene usually has strong vertical and horizontal edges, whereas coast scene tends to have edge randomly distribution various direction. So we define global features in an image in terms of an edge direction histogram. A total of 73 bins are used to represent edge direction histogram of an image; the first 72 bins and the last are used to represent edge direction quantized at $5^\circ$ interval and the sum of the pixels number that do not contribute to any edge, respectively. The edge direction histogram is defined as

$$h(d) = (h_i(d))_{i=1..73}$$

where $h_i(d)$ is the count of the edge direction in the $i$th bin in the histogram.

2.3 Selecting Weight

Our purpose is to select the weights which can make the extension representation more powerful according to the content of the image. For example, if the entire image is full of the interest points, the BOV can acquire the nature of image easily. So, the weight value will be larger under the situation. While the interest points centralize in some small parts of image, the image representation needs more global information to classify the scene better, which means that the weight value is small under this condition and global information play a little more important role in representation of an image.

Now we move on to the problem of selecting weight. In order to create a more thorough image representation, we extend the standard BOV methodology by weighted concatenation global features.

Here, we assume that the two types of features (local and global) are independent, and can be fused by weighted concatenating. The weight indicates the relative importance (versus global) of local information described over the interest points in the scene classification, so it is defined according to the distribution of interest points in the image. To achieve this goal, an image is divided into several rectangular regions, see Fig. 2. In particular, the image is first partitioned into $K \times K$ sub-blocks and then the histogram is

Fig. 1 A general framework of the classification using EBOV scene representation.

Fig. 2 Selecting weight scheme.
built based on the amount of interest points in each block. We use such histogram to characterize the distribution of local information. The entropy of histogram illustrates the distribution degree of local information in an image.

\[ H_d(x) = f(p_0, p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2(p_i) \]  

(3)

where \( p_i \) denotes the number of \( i \)-th bin in the histogram. We quantize entropy \( H_d(x) \) into weight \( w_i \) and all \( w_i \) comprise the set \( W \).

\[ w_d = \frac{H_d(x)}{\max_{d \in D} H_d(x)} \quad w_d \in W \]  

(4)

where \( D \) denotes the set of all images in the dataset. The interesting points reflect the invariability rather than the gist of image, and the invariability may remove some information that is valuable for classification. To solve this problem, we must take global information in account. To integrate the BOV feature \( l_d \) and global information \( g_d \) with optimal weight \( w_d \) according to:

\[ H_{a+b}(d) = [w_d \times l_d, (1 - w_d) \times g_d] \]  

(5)

Where \( H_{a+b}(d) \) denote the new scene representation of the image \( d \).

2.4 pLSA Model

The result of classification can be acquired with the help of pLSA algorithm by introducing the new extension representation. Probabilistic Latent Semantic Analysis (pLSA) is a generative model from the statistical text literature [8]. In text analysis the pLSA model is used to discover topics in a document using the bag-of-words representation. Here we have images as documents and we discover topics as object categories (e.g. grass, houses), so that an image containing instances of several objects is modelled as a mixture of topics. This topics distribution over the images is used to classify an image belonging to a certain scene [1]. A joint probability model is defined:

\[ p(w/d) = \sum_{z \in Z} p(w/z)p(z/d) \]  

(6)

where \( p(w/d) \) denotes the image \( d \) representation, and it can be replaced by \( H_{a+b}(d) \) which was computed above in our experiment. Each image is modeled as a mixture of topics

\[ p(z/d) ; \ p(w/z) \] represent the probability of observing the vistern given the aspect [4].

The categorization which predicts the classes of unlabeled images using pLSA model can be divided into two separate steps: (as Fig. 3): training and testing. During training, the parameter \( p(w/z) \) is learnt from the set of training images. Each training image is then represented by a Z-vector \( p(z/d_{train}) \), where \( Z \) is the number of topic learnt. Testing stage computes specific mixing coefficients \( p(z/d_{test}) \) using the fold-in heuristic model. The test image is then labeled using a K nearest neighbors scheme (KNN) on \( p(z/d_{train}) \) of the training images

3. Experiment

3.1 Datasets

In our experiment we use the dataset in [6]. Each scene category is randomly divided into two separate image sets, 50 for training and 50 for testing, respectively. A confusion table is used to illustrate the performance of the approach. In the confusion table, the x-axis represents the result of the proposed approach for each scene category. The y-axis represents the ground truth categories of scenes. The orders of the scene categories are the same in both axes. Hence in the ideal case one should expect diagonal lines include all the testing data which show perfect discrimination power of the category models over all categories of scenes.

4. Results

Through Fig. 4, we take a close look at the influence and interaction of the local and global information in scene classification. Different images have different weights, even in the same category. Table 1 is an overview of performance of this approach. The confusion table reveals that the highest error rate occurs in the mountain and coast categories.
Table 1 Confusion table using 50 training and 50 testing images from each category. The average performance is 90%.

<table>
<thead>
<tr>
<th>category</th>
<th>coa.</th>
<th>for.</th>
<th>hig.</th>
<th>city</th>
<th>mou.</th>
<th>cou.</th>
<th>str.</th>
<th>bui.</th>
</tr>
</thead>
<tbody>
<tr>
<td>coa.</td>
<td>42</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>for.</td>
<td>0</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>hig.</td>
<td>0</td>
<td>1</td>
<td>48</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>city</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>48</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>mou.</td>
<td>3</td>
<td>0</td>
<td>6</td>
<td>40</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>cou.</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>45</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>str.</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>47</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>bui.</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>44</td>
</tr>
</tbody>
</table>

Table 2 Classification result comparison (1: BOV; 2: EBOV).

<table>
<thead>
<tr>
<th>dataset</th>
<th>#of categor.</th>
<th>Anna perf.</th>
<th>Our perf.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oliva</td>
<td>8</td>
<td>87%</td>
<td>90%</td>
</tr>
<tr>
<td>4-natural</td>
<td>90.2%</td>
<td>92%</td>
<td></td>
</tr>
<tr>
<td>4-manmade</td>
<td>90.5%</td>
<td>93%</td>
<td></td>
</tr>
</tbody>
</table>

And the highest correct rate occurs in the forest and street categories.

Table 2 shows how the global information influences the result. The largest improvement occurs in the country scene. Because abundant homogenous image regions are ignored by interest points detectors in this scene, the local invariant information is not enough for discrimination. Two scenes, city and street, which are indistinguishable from each other when analyzed by the conventional BOV, might be discriminated into correct class with help of global information. The reason is that the local invariant information for both scenes is similar. Thus it can be seen that combining the global information with local information can improve the discriminative ability.

We compare our classification approach performance with the previous approach [1], [6] on Oliva [6] dataset in the Table 3. It is obviously that our approach is better than the approach of Anna [1]. The results show that global information can overcome the shortcoming of local information which only describes the interest points that generally occur on the richer texture parts of image. However, there are several homogeneous regions in the image that can’t detect the interest points but represent the image scene. So the local information can not fully represent the categories of images. The gist of these images (as the right column of the first row in Fig. 4 shows) should be obtained by more global information. On the other way some images, such as left column of the first row in Fig. 4, can be represented very well by local information. Considering the results presented in this paper, we believe that the presented extension to the standard BOV representation is an effective scene modeling methodology.

5. Conclusion

In this paper, we propose a framework to construct a novel scene representation using the local and global information. It is obvious that categorization performance using either local or global information separately is lower than when integrating these two kinds of image information together. However, the task of scene classification still leaves room for improvement. This is a promising direction, and we might be able to form much more powerful representation for difficult categories.

Acknowledgment

This work was supported by the Chinese National Programs for HTRD (Project No.2007AA01Z168).

References