報告番号	※甲　　第　　号

主　論　文　の　要　旨

論文題目　A Study of Informative Patch Extraction and Patch-level Context Exploration for Image Categorization（画像分類のための情報豊かなパッチの抽出とパッチレベル文脈探索の研究）

氏　名　白　双

論 文 内 容 の 要 旨

Nowadays, huge numbers of digital images are produced every day. It has become urgent to organize and access them efficiently. Therefore, image categorization, which means to classify images into general categories based on their contents, attracts more and more attention. In order to construct a qualified image categorization system, there are some problems to cope with, such as view and lighting change, occlusion and background clutter as well as intra-class dissimilarity and inter-class similarity. All these problems are typical in the real world images.

At first, in order to create a representation of an input image, global features such as color and texture are extracted and utilized. Although methods using global features give some achievements, they cannot give satisfactory performance for practical applications due to the problems mentioned above. In contrast, approaches based on local features have shown their superiority over global features. In local feature based approaches, image representations are created from local image patches which are extracted from input images through interest point detectors or dense sampling and described by patch descriptors such as SIFT.

Currently, among various local feature based approaches, a method called bag of visual words gives state of the art results. In the framework of this method, the first step is to create a codebook of visual words by applying vector quantization to local image features obtained from training images. Next, given an input image, local image patches are extracted from it and described by a certain patch descriptor. Thereafter, each obtained local feature is encoded by its nearest visual word of the codebook in Euclidean space. As a result, the input image is represented as a histogram indicating the frequency of each visual word appearing in it. This procedure enables us to produce robust and characteristic image representations, which can subsequently be classified by using off-the-shelf classifiers.

The simplicity and effectiveness of the bag of visual words method has made it one of the most promising approaches to image categorization. However, there are still problems to cope with. For instance, patches extracted on the basis of commonly used patch sampling strategies include too much noise. At the same time, in most of the previous works, local image features are handled independently, which results in considerable information loss. Therefore, in order to create discriminative image representations, approaches to overcome these disadvantages are needed.

The purpose of this thesis is to investigate the two above-mentioned issues of the bag of visual words method: informative patch extraction and patch-level context exploration, which influence the final image categorization performance significantly.

The first issue is on informative patch extraction. Until now, the commonly used image patch extraction methods can be divided into two main branches. The first branch is based on dense sampling or interest point detectors. These methods are not able to extract informative image patches for the task of image categorization. The other branch is through exhaustive sampling together with a feature selection process. It is expensive in computation and memory usage. Therefore, to cope with this problem, we focus on designing an effective image patch extraction approach.

In the proposed method, both bottom-up processing and top-down processing are utilized for evaluating image regions with respect to their informativeness for image categorization. Thereafter, corresponding to each kind of processing, a separate saliency map is created, which is then combined with each other in a weighted manner. Finally, we perform image patch extraction based on the combined saliency map. The proposed approach can use informative image patches to create image representations, at the same time, no resources of computation and memory are wasted for extracting and evaluating useless raw patches.

The second issue is about patch-level context exploration. In order to create discriminative image representations, we try to explore patch-level context to incorporate relationships among patches. Although some works on patch-level context have already existed, in these works the combination of patches of interest and their context patches is fixed. Variations of the relationships among the patches are not taken into consideration. In this thesis, we propose an approach in which the patch of interest is flexibly combined with its context patches.

In this approach, a patch of interest is associated with other patches, such as patches extracted from the same sample point but of different scales or patches in its neighborhood. When creating image representations, the associated patches are taken as a patch set and dealt with together. Specifically, for encoding a patch set, each of its patches is encoded independently first. After that, the encoding of the patch of interest is adjusted based on its context patches. So that their relationships can be utilized.

Finally, for each proposed method, we conducted extensive experiments to assess its effectiveness. We evaluated the patch extraction method using an object category dataset and a scene category dataset and evaluated the patch-level context exploration method using three scene category datasets. We compared the proposed method to other works. Obtained results have demonstrated the effectiveness of the proposed methods.

The thesis is organized as follows.

Chapter 1 is the introduction of the thesis. It gives the research background and objectives, as well as the thesis characteristics and overview.

Chapter 2 presents a brief overview of previous research which has utilized feature selection or context modeling. The originalities of the proposed approaches compared with the related conventional methods are also described.

Chapter 3 explains some of the basic methods that are used in this thesis. Specifically, we described the bag of visual words (BOVW) framework, the scale-invariant feature transform (SIFT) descriptor, the Gaussian mixture model (GMM) and the support vector machine (SVM) classifier. The bag of visual words framework is adopted for creating image representations from local features. The scale-invariant feature transform descriptor is employed for obtaining local image features from extracted image patches. The Gaussian mixture model is used to model the interest point distribution. Finally, the support vector machine classifier is used to classify obtained image representations.

Chapter 4 proposes a saliency based image patch extraction strategy. In the proposed method, given a regularly divided input image, each of its grids is evaluated based on both its low-level image information and the statistical information of the training images with respect to its informativeness for image categorization. For evaluating an image grid on the basis of the low-level image information, we first detect interest points in the input image. Then, the interest point distribution is investigated and modeled as Gaussian mixture model. After that, the interest point distribution model obtained is used for assessing image grids. Based on the image grid assessment, a saliency map is created. On the other hand, the statistical information of the training images is employed as an image grid informativeness measure. In this procedure, first a set of clusters are created by applying k-means to training image grid features. Thereafter, the mean vector, covariance matrix and entropy of each created cluster are calculated and recorded. Finally, the informativeness of an image grid is determined by its distance to the grid clusters and the calculated statistical information. After the informativeness of every grid in the input image has been measured, another saliency map is created. When these two saliency maps are obtained, we propose a weighted fusion method to combine them. In the last stage, image patch extraction is performed based on the combined image saliency map.

To assess the proposed method, we describe the extracted patches by SIFT features and use them to create image representations on the basis of the bag of visual words method. An object category dataset and a scene category dataset are used for experiments, respectively. Obtained results demonstrated that the low-level image information is more effective for the object category dataset, while the statistical information of the training image set gains better performance for the scene category dataset. Furthermore, results by combining these two sources of information show superior performance than all methods used for comparison.

Chapter 5 focuses on incorporating patch-level context into image representations. In this chapter, to relieve ambiguity existing in the process of image patch encoding, a

patch of interest is associated to other patches such as patches extracted from the same sample point but of different scales or patches in its neighborhood. The associated patches are taken as a patch set and dealt with together. Given an input image, multi-scales of patches are extracted from it and associated to one another. To create its image representation, each patch is encoded independently first. After that, the encoding of the patch of interest is adjusted based on its context patches. Consequently, in the obtained image representation, relationships among patches are incorporated. Furthermore, in order to explore image information extensively, for an input image, we create three representations based on different context strategies, i.e. image representation without context, image representation with different scale patch context and image representation with nearby patch context. The obtained image representations are fused in the image classification stage in a probabilistic manner to predicate the input image label.

We used the SVM classifier for performing the image representation classification. Datasets scene categories 8, scene categories 13 and scene categories 15 are used to evaluate the proposed method. Obtained results demonstrated its effectiveness.

Chapter 6 summarizes this thesis and gives some ideas for future research.

学位関係