Acta Automatica Sinica
Vol. 40, No. 4 April, 2014

Data Driven Hierarchical Serial Scene Classification Framework

FENG Wen-Gang[1, 2]

Abstract Scene classification is a complicated task, because it includes much content and it is difficult to capture its distribution. A novel hierarchical serial scene classification framework is presented in this paper. At first, we use hierarchical feature to present both the global scene and local patches containing specific objects. Hierarchy is presented by space pyramid match, and our own codebook is built by two different types of words. Secondly, we train the visual words by generative and discriminative methods respectively based on space pyramid match, which could obtain the local patch labels efficiently. Then, we use a neural network to simulate the human decision process, which leads to the final scene category from local labels. Experiments show that the hierarchical serial scene image representation and classification model obtains superior results with respect to accuracy.

Key words Space pyramid match, visual codebook, generative method, discriminative method, neural network

Citation Feng Wen-Gang. Data driven hierarchical serial scene classification framework. *Acta Automatica Sinica*, 2014, **40**(4): 763−770

Nowadays, as it becomes increasingly viable to capture, store, and share large amounts of image and video data, automatic image analysis is crucial to managing visual information. The amounts of the image data and the number of the users have an exponential increase. Image scene classification, which can be used for indexing, searching, filtering and mining large amounts of image data, becomes increasingly important for users. For example, we can group the images according to the high level concepts that contain or the scenes that occur in such as "office", "mountain", "street", and so on, for efficient image searching. So we need a robust and efficient image scene classification system. The goal of machine learning is to build automated systems that can classify and recognize complex patterns in data.

Scene classification[1−3] plays an important role in efficient image resource management and automatic image classification, and it is very important in computer vision and pattern recognition domain as well as content-based retrieval. The goal of a computer vision system is to build a model of the real world that allows a user to interact with it. The ultimate goal is to allow a computer to "see" the world in a manner similar to the biological visual system. The goal of this research is to allow images to be quickly, accurately, and efficiently classified based on the image semantics and relevance.

Humans can recognize one scene in a very short time because scene detailed information is not considered in rapid serial visual presentation (RSVP) task. The model needed to classify a scene has been trained sufficiently to fit most situations. Based on RSVP, we propose a novel hierarchical serial scene classification framework that has superior results for scene recognition.

1) Related work

Scene classification work could be considered as the mission hard to be completed, because even one scene still includes many contents in it. But why RSVP could recognize object in short time, because it ignores details. Most scene classification work can be divided into two encampments local feature based scene classification and global feature based scene classification. Both of the methods could recognize scene efficiently, and it is opined that the two methods would be complementary to each other.

Global feature based scene classification[4−11], which compactly summarizes the image′s statistics and semantics, identifies whole scenes, not small patches of objects or precise region boundaries within the scenes. The challenge of discovering a compact and holistic representation for unconstrained images has hence prompted significant recent research. Oliva and Torralba[5] introduce a representation of the structure of real-world scenes termed "spatial envelope", for which properties provide a holistic description of the scene without local object information.

Local feature[12−14] based scene classification[15−21] (i.e., recognizing the whole image based on every patch) uses segmented image regions and their configuration relationships to recognize the scene. The bag-of-words model has been used in scene classification domain quite extensively in recent years.

2) Our approach

In this framework, first of all, we extract hierarchical feature by space image pyramid with which global and local representations can be surprisingly effective, what is valid for identifying the overall scene. It is also effective for categorizing images as containing specific objects, even when these objects are embedded in heavy clutter and vary significantly in pose and appearance. We use rapid serial visual presentation which has degraded to visual information processing to express one image, which includes the global and local manners, and build our own codebooks by two different types of features.

Secondly, we train the visual words by generative and discriminative methods respectively. Because one image would be divided into three layers and 21 patches, whose feature extraction and model training are independent. K-nearest neighbor (KNN) algorithm is used to obtain the single-estimate probability and label for each patch.

Finally, we use a neural network which would obtain the final scene category label from all the local labels, what makes the decision like human beings.

3) Organization of the rest of the paper

The rest of this paper is organized as follows. Section 1 outlines the proposed framework with various components such as hierarchical feature extraction, generative model, and discriminative model described in various subsections. Experimental results are presented in Section 2 followed by

Manuscript received June 15, 2012; accepted July 19, 2013
Recommended by Associate Editor TIAN Jie

1. Department of Policing Intelligence, Chinese People′s Public Security University, Beijing 100038, China 2. Public Security Intelligence Research Center, Chinese People′s Public Security University, Beijing 100038, China

the conclusions in Section 3.

1 Proposed scene classification framework

We first describe the hierarchical serial scene classification framework in general, and then introduce each part of the framework in detail.

1.1 Hierarchical serial scene classification framework

In the training step, there are five stages: First, each one image will be presented as 21 images (patches) in three layers using an image pyramid. Second, we will build codebooks, which have a clustering part (SIFT) and non-clustering part (color, texture, gist, and fractal features). Third, each visual word extracted from one image patch is trained in a generative way, probability latent semantic analysis (pLSA) model. Each training patch is then represented by a z-vector ($p(z|d)$), where z is a topic and d is a document (image), and the number of the z vector is the number of the topic learned. Fourth, the z-vectors (patch label) obtained from last step are subjected to discriminative training (KNN model). Finally, we use a neural network to get the final one classification category depending on the 21 patches scene categories. This approach simulates human′s decision process based on local and global visual perception.

The test step also proceeds in five stages. The difference is that the test image is projected onto the simplex spanned by the vector learned during training. This is achieved by running expectation-maximization (EM) algorithm in a similar manner to that used in learning, but now, only the z-vectors are updated in each M-step with the other learned vectors ($p(w|z)$) are fixed. The result is that the test image is represented by a z-vector. The test image is then classified by the multiclass discriminative classifier. Fig. 1 shows graphically the hierarchical serial framework for both the training and testing steps.

1.2 Hierarchical feature

1.2.1 Space pyramid match

The human′s observation of the scene and object is a hierarchical model moving from global to local. According to the pyramid data structure[1, 22], the image could be defined as a sequence $D = \{D_k\}_{k=1}^{K}$, and K is the depth which is the key point observed by a human. On this basis, the image is further divided into a planar sequence, $D_k = \{D_i^k\}_{i=1}^{4^{i+1}}$, where D_i^k is the image content of the subscript area R_i^k in the image D.

$$D_i^k = \{D_{x,y}, (x,y) \in R\}_{i=1}^{4^{i+1}} \tag{1}$$

where R is the subscript matrix of the image D, and R_i^k is the subscript set of the i-th patch of the k-th layer in the pyramid. Here we use three layers as shown in Fig. 2.

$$R = \{R_i^k\}_{i=1}^{4^{i+1}},\ R = \{(x,y)|x = 1, \cdots, X,\ y = 1, \cdots, Y\} \tag{2}$$

where (X, Y) is the pixel resolution of image D. The image scale space maintains the entropy structure of the image. This is because there is no filter processing on the original image and every layer of an image region description is coherent with the constraint of scale consistency.

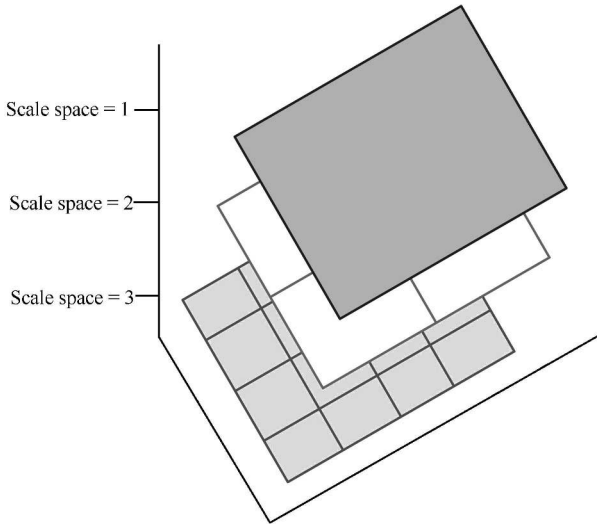

Fig. 2 Image scale space

The image pyramid is constructed by placing a sequence of increasingly detailer grids over the image space. Here we use a three-layer pyramid: The first layer is the original image, which means one patch; the second layer has four grids, which means each patch is a quarter of the original image; the last layer has sixteen grids, which is one sixteenth of the original image. Thus we have 21 images (patches) from the original image.

1.2.2 Feature extraction

In the framework, first an input image is separated into three layers and 21 patches by an image pyramid. We treat each patch from the original image as one independent image. That means that the scene dataset will have 21 subsets. This is the parallel property as shown in Fig. 3.

For each patch we extract two types of features to build codebooks used in the experiments of Section 2. First, we have the so-called "clustering codebook", which are SIFT (scale-invariant feature transformer) descriptors of 16×16 pixel patches computed over a grid with spacing of eight pixels. We perform K-means clustering from the training set to form a visual codebook.

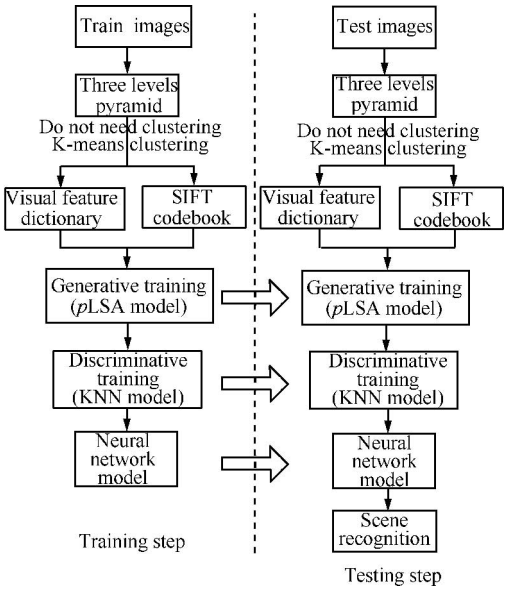

Fig. 1 Flowchart of the algorithm