The Automatic image annotation for scenery images based on multi-class support vector machines

THE AUTOMATIC IMAGE ANNOTATION FOR SCENERY IMAGES BASED ON MULTI-CLASS SUPPORT VECTOR MACHINES

Yanyu Gao*, Xinping Wang(D1), Takashi Uozumi**
* SVBL ** Department of Computer Science and Systems Engineering, Muroran Institute of Technology

1. Introduction

As low cost digital cameras have become available and Internet and multimedia information technology are being rapidly developed, huge amounts of non-textual photographs in electronic form are available. Manually annotating so many digital images is tedious and prohibitively expensive. Automatic annotation using computers and image understanding technique can undoubtedly reduce cost and save labours, but its annotation precision and flexibility are inferior to manual annotation. How to improve precision and flexibility of automatic annotation to make it applicable to common photos has attracted our interests for a long time.

In recent years, a variety of image auto-annotation systems have been proposed due to the development of artificial intelligence and statistical learning theory. According to processing objects of feature extraction and annotation, auto-annotation models can be classified into three classes: 1) image-based auto-annotation [1] that regards the whole image as an individual visual pattern and uses visual features of the whole image to infer its semantic contents; 2) blob-based (region-based) auto-annotation [2,3] that takes the homogeneous image region or connected homogeneous image regions with the same visual attributes as the annotating object and extracts its visual features for blob understanding; 3) salient-based auto-annotation [4] that regards the salient regions as annotating objects and extracts their visual features for image understanding. Among the three kinds of annotation models, blob-based auto-annotation received more attention. One of its first attempts was reported by Mori et al. [3], who calculated the co-occurrence between annotation words and image regions created by a regular grid, and applied the probability to predict image contents.

Although research in automatic image annotation has made great efforts in improving annotation precision and speed as well as enlarging the scope of annotation objects, at least three issues need to be given more attention henceforth.

1) Recognizing scenery objects by image analysis is easily affected by impersonal elements, such as lighting conditions, as well as subjective elements, such as photographing angles.
2) So far, image segmentation still remains to be fragile and error-prone and has not obtained perfect results for the simple reason that segmentation itself depends on the output of interpretation [5].
3) Different people may give different labels to the same scenery object because of the lingual diversities and cognition differences.

In order to solve these problems, we make great efforts in keywords selection, feature extraction, and region annotation. We propose to represent each region with a set of color and texture features that is insensitive to size, orientation and shape of the region. To make the annotation keywords acceptable to most people, a questionnaire experiment is performed, by which the uncommonly used words are removed. To label each region exactly and clearly, we adopt the directed acyclic graph SVM (DAGSVM), which has good generalization ability and can complete multi-classification in short time. To improve annotation precision furthermore, we define a set of logical rules based on the image context and spatial constraints and use them to revise improper labels.

The paper is organized as follows. Section 2 introduces the system framework and pre-processing. The multi-classification method by DAGSVM is explained in Section 3. Methods for correcting improper annotations are explained in Section 4. Experiments and results are discussed in Section 5. Section 6 gives the conclusion and future work.

2. System Framework and Preprocessing

2.1 System framework

Our system is a blob-based auto-annotation system, which consists of 4 steps (as shown in Fig.1): 1) Segment: the input image is segmented into multiple regions so that each region corresponds to a single object. 2) Feature extraction: extract color and texture features for each region skillfully, so that they can effectively decrease the influence of inconsistent image scales and various photographing angles. 3) Annotation: annotate each segmented region by the multi-class classifier DAGSVM as well as the supervised learning method. 4) Correction: define a set of logical rules based on the image context and spatial constraints and use them to revise improper annotations.

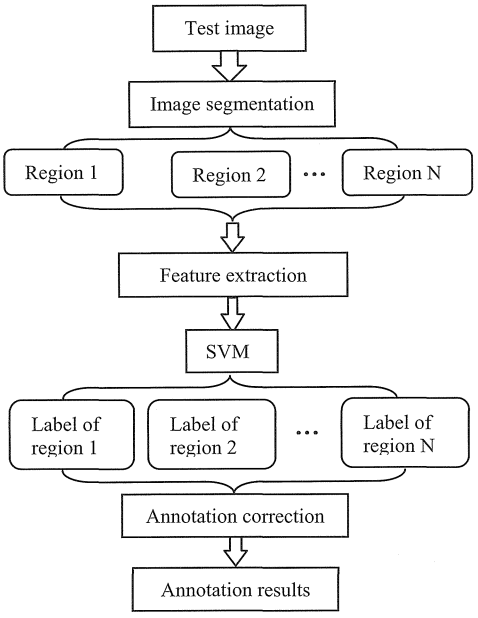

Fig.1 The basic framework of image auto-annotation system

2.2 Semantic keywords selection

In this paper, we focus on annotating natural scenery images. For reducing the lingual diversities and cognition differences, a small questionnaire experiment is performed to help us select the semantic keywords. 4 subjects (2 male and 2 female) with normal color vision and normal or corrected-to-normal vision attended the experiment. They were asked to watch 300 carefully selected natural scenery images and name each object in those images. Those images include various scenery objects in different lighting conditions photographed from different angles. After deleted those rarely used synonymies, we group these keywords into two hierarchies (as shown in Table 1). Some words are very general and are used to describe a large range of scenery objects (e.g. stone), while others are only fit for describing a specific object (e.g. pebble).

Table 1. Semantic keywords organized in 2 hierarchies

General keywords	Detailed keywords
Water	Sea, river, lake, waterfall, etc.
Stone	Pebble, cobblestone, gravestone, stonewall, reef, etc.
Soil	Soil, sandy soil, sandbeach, desert, etc.
Mountain	Hill, ice-mountain, cliff, island, snow-mountain, etc.
Sky	Clear sky, dark sky, sky with sun, sky with white clouds, sky with dark clouds, sky with red clouds, etc.
Herbaceous plant	Winter grass, green grass, crop, etc.
Woody plant	Bush, big tree, small tree, reed, weed, etc.
Flower	Red flower, yellow flower, creeper, etc.
Road	Alley, aisle, lane, channel, pavement, railway, etc.
Building	Arena, temple, castle, brick building, wood building, fence, sculpture, etc.
Vehicle	Car, airplane, balloon, truck, ship, etc.
Animals	Human, tiger, dolphin, elephant, horse, bear, etc.

2.3 Image segmentation

In our system, pixel-clustering based spatially constrained mixture model [6] is adopted for image segmentation, which considered the pixel location information and tried to assign the same cluster label to spatially adjacent pixels. In order to reduce the segmentation time, we restrict the original iteration cycles to 20. Through some simple tests, we found that to most of scenery images, 20 runs engender little change to segmentation results comparing to those by 50 runs. The falsely segmented regions will be revised in the step of annotation correction.

2.4 Visual Feature Extraction

In order to represent each region effectively, we propose a feature combination scheme, which describes each region with a set of feature vectors that consist of color probability distribution [7] and Gabor wavelet texture features [8].

1) Color probability distribution

Color feature needs to be defined in a selected color space. In this paper, the CIE $L^*a^*b^*$ color space is adopted since it is regarded as the most complete and perceptually uniform color space in the sense that the differences between points plotted in the color space correspond to visual differences between the colors. Color probability distribution can be uniquely characterized by the first, second and third-order central moments of the color distribution in a region. The first moment calculates the average color values in color channels L, a, b. The second and the third central moment reflect the variance and the skewness of each color channel. Since all color central moments have the same units, it is easy to realize similarity measurement.

2) Gabor wavelet texture features

Gabor wavelet texture has been proved to be an effective texture feature, whose basic idea is to extract features at multiple scales and orientations using the Gabor wavelet decomposition. In this paper, the mean and standard deviation of the magnitude of the Gabor transform coefficients in three scales at four orientations are calculated that constitute 12 feature vectors, each of which has two elements.

We combine each texture feature vector and 9 color features to form a visual feature vector. Totally 12 feature vectors can be extracted to represent an image, where each vector includes 11 elements (2 texture features and 9 color features). For the training sample, all of the 12 feature vectors are listed as the reference standard, which reflect the scenery object from different directions and in difference scales. For the test image, we randomly select one feature vector to represent a region and compare it with standard vectors. Such kind of

feature representation can effectively decrease the influence of inconsistent image scales and various photographing angles. To eliminate the impact of inconsistent feature ranges, we normalize each feature component to be a variable of zero mean and unit variance by $F=(f_i-\mu_i)/\sigma_i$, where μ_i and σ_i are the mean and standard deviation of feature f_i over the entire image database.

3 Annotation by SVM

After segmentation, we hope to label the segmented regions $S=\{s_1, s_2, \ldots, s_i\}$ with a set of semantic keywords $\{g_1, g_2, \ldots, g_i\}$ as accurately as possible. The annotation process can be viewed as a kind of pattern classification, where labeling each region is equal to classifying each region into one of the predefined semantic classes. In order to obtain a high recognition precision, supervised machine learning method is recommended, which constructs the relationship between semantic concept and visual features by learning a set of predefined image examples.

SVM is a popular and effective classifier that adopts the supervised learning method for pattern classification. Based on the principle of structural risk minimization, SVMs have better generalization performance than other traditional classifiers (i.e., neural networks), and can yield high recognition accuracy with small training sets [9].

However, SVMs were originally designed for two-class problems. To make them adaptable to real applications, three kinds of multiclass classification techniques have been proposed, namely one-against-all, one-against-one, and directed acyclic graph SVM (DAGSVM) [10]. The one-against-all approach constructs L binary SVM classifiers, each of which is trained to separate one class from the rest (L-1) classes. Here L is the total number of classes. The one-against-one method trains L(L-1)/2 binary classifiers with each classifier separating a pair of classes.

In this paper, we adopt the DAGSVM to classify each region, which needs less testing time than the one-against-one SVM and has better generalization ability than the one-against-all SVM. In the training phase, DAGSVM builds L(L-1)/2 binary SVMs with each SVM separating a pair of classes. In the test phase, a rooted binary directed acyclic graph is constructed, which has L(L-1)/2 internal nodes and L leaves (as shown in Fig.2). Each internal node is a binary SVM that distinguishes two classes. DAGSVM first initializes a list that includes all classes. Then, at the root node, the test region s_i is evaluated against the decision node that corresponds to the first and the last elements of the list. If the node prefers one of the two classes, the other class is eliminated from the list, and the DAGSVM proceeds to test the first and the last elements of the new list. Such evaluation proceeds until only one element remains in the list. The leaf node indicates the predicted class that the test region s_i belongs to.

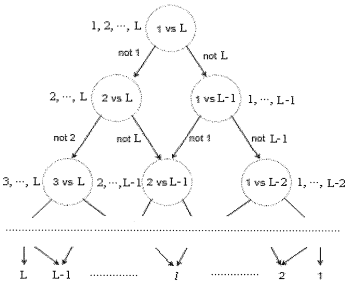

Fig.2 The working flow of DAGSVM for multiclass classification

4 Adjusting annotation by contextual relationship and region attributes

Because of many subjective and objective factors, such as segmentation error, unusual lighting conditions, and similar appearance of different scenery objects, recognizing isolated regions only with color and texture features is error-prone, although these features have high representative ability and the classifier has better differentiability and generalizability. In order to improve the annotation accuracy, coexistence, relative location, and circularity are investigated.

Coexistence is used to judge whether two objects could coexist in an image. For example, sea is more often associated with sand beach, sky, boat, or cliff and less often with objects like tiger and desert. Relative location is used to judge whether the annotation of a segmented region is rational according to its relative location to other objects. For example, sea and sky are easily mixed because they sometimes have similar color and texture features. However, if the upper part of an image were annotated as sea and the lower part were annotated as sky, in common sense we regard the annotations illogical. Region circularity is mainly used to roughly evaluate the shape of a region. By comparing the region shape and label, we can find some inaccurate annotations. For example, a brick building would be misunderstood as sun, because sometimes they have similar color features. Considering that sun is usually in circular shape, while a brick building is in quadrate shape, circularity is a good tool of judgment.

4. Experiments

Our image set consists of 3000 scenery images selected from the Corel Stock Photo Library, which involve various contents and themes, such as fields, waterfalls, sunrises and sunsets, coasts, and deserts. We manually cropped a set of single-object region from 1000 carefully selected images and use them as our training samples, which involve at least 10 training samples for each fine semantic classes listed in Table 1. The visual features of all training samples are calculated and constitute the standard feature database. The rest 2000 images are used to test and evaluate our system.

Every test image is first segmented by spatially constrained mixture model. The number of segments may be different from image to image. We set the number of segments as c=6 initially and decrease the number if a region is smaller than 2.5% of the image's area. Then all test images are labeled with rough keywords and fine keywords after DAGSVM classification and logical correction. Two annotation examples are shown in Table 2.

Table 2. Comparison between manual annotation and auto-annotation

Images		Rough annotation	Fine annotation
	A	Mountain, sky, trees, grass	Snow mountain, clear sky, winter trees, bush
	M	Mountain, sky, trees, human	Snow mountain, sky with white clouds, winter trees, man
	A	Water, stone, sky	Sea, stonewall, dark sky
	M	Water, mountain, sky	Sea, reef, clear sky

PS: 'A' means auto-annotation, 'M' means manual annotation

From Table 2, we found that sometimes our system would confuse tree with grass, stone with mountain, and sometimes it regards mountain covered with trees as trees only. Although spatial relationship can revise some errors, lots of errors cannot be discovered yet.

We also use the retrieval precision, calculated by $P_r=N_c/N_r$, to evaluate the annotation results, where N_c represents the number of retrieved images that contain query keyword in their annotation, N_r is the number of retrieved images. 5 rough keywords—water, sky, mountain, vehicle, soil; and 5 fine keywords—sea, clear-sky, snow-mountain, bus, and desert are tested. Their precisions are shown in Fig.3.

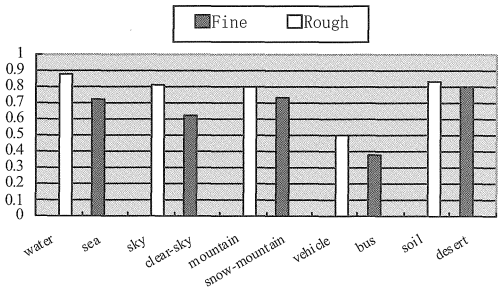

Fig. 3 Annotation precision of rough annotation and fine annotation

From Fig.3 we found that: 1) DAGSVM can obtain a high recognition precision for sky, sea, mountain, and soil, but a low precision for vehicle. It seems that color and texture features have enough representative for the former objects, but less representative for the latter object. Although the exact shape feature is very important to vehicle recognition, constructing a uniform shape description for all kinds of vehicles is too difficult to realize. 2) Usually, annotation with rough keywords can achieve a higher precision than that with fine keywords. One possible reason is that fine semantic classes in a rough semantic class have more similarities.

5. Conclusion and future work

Automatic image annotation has been investigated for many years. In this paper, we propose an auto-annotation system that consists of four parts: keywords selection, feature extraction, region annotation, and annotation correction. Experiments performed on a small training set got an encouraging result. Our current annotation objects are limited to outdoor scenery images selected from a certain image database. In the future, we would focus on annotating broad-content images and consider applying the ontology technique to organize the semantic keywords.

References

[1] A Yavlinsky, E Schofield and S Rüger, "Automated image annotation using global features and robust nonparametric density estimation," *Int'l Conf on Image and Video Retrieval (CIVR)*, Singapore (2005).

[2] P Duygulu, K Barnard, J de Freitas, and D Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," 7th *European Conf. Computer Vision*, pp.97–112, Copenhagen, Denmark (2002).

[3] Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," *The First Intl. Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM'99)*.

[4] J. Tang, J. S. Hare, and P. H. Lewis, "Image Auto-annotation using a statistical model with salient regions," *Proc. IEEE Int'l Conf. Multimedia & Expo (ICME)*, Toronto, Canada (2006).

[5] K. Sunil Kumar, U. B. Desai, "Joint segmentation and image interpretation, " *Pattern Recognition*, vol.32, pp. 577-289 (1999).

[6] K. Blekas, A. Likas, N. Galatsanos and I. Lagaris, "A Spatially-Constrained Mixture Model for Image Segmentation," *IEEE Trans. Neural Networks,* vol.16, no.2, pp.494-498 (2005).

[7] M. A. Stricker and M. Orengo, "Similarity of color images", *Proc. SPIE Storage Retrieval Still Image Video Databases IV*, vol. 2420, pp.381-392 (1996).

[8] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans.PAMI,* vol.18, no.8, pp.837-842 (1996).

[9] V. Vapnik, *Statistical Learning Theory*, Wiley, New York (1998).

[10] J. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin DAGSVM's for multiclass classification," in *Advances in Neural Information Processing System,* Cambridge, MA: MIT Press, vol.12, pp.547-553 (2000).