ISSN: 0976-9102 (ONLINE)
DOI: 10.21917/ijivp.2018.0250

SEMANTIC IMAGE DESCRIPTION AND CLASSIFICATION BASED ON GENERALIZED SET

Ri ChangYong, Pak DuHo, Rim KyongChol and Ju JinHok
Institute of Information Science, Kim Il Sung University, D.P.R. of Korea

Abstract

A semantic image description model based on generalized set is proposed, and the semantic similarity (distance) measure between images is presented. Semantic image information can be completely represented in this model as compared with previous researches based on vector space. The semantic image description model based on generalized set is similar to human understanding of image knowledge. For the purpose of the semantic image classification, semantic distance based on support vector machine classifier is employed. Experimental results show the validity of new method, and that the image classification accuracy is improved.

Keywords:

Semantic Image Description, Image Classification, Generalized Set

1. INTRODUCTION

Semantic image description is the representation of knowledge for contents what the images mean. For semantic image classification, we should make suitable image knowledge representation and take knowledge contained in image. The knowledge representation schemes have generally both ways, Sanders et al. [10]: knowledge representation based on feature and structure. The way based on feature describes a pairs of "property-value", a feature vector or a framework. The way based on structure describes the internal and structural relationship of objects. Human perception for image contents is based on the objects contained in an image and the relationship between objects. Describing the image contents using only the simple features (type of numeral value, vector or matrix), we could not process the semantic information of image completely.

Recently, many researchers studied the semantic image description methods and its applications. Nwogu et al. [7] discussed the semantic image description based on ontology. They proposed the image description model based on image grammar and used an ontological tree with the objects and relationship between objects. Boutell et al. [2] supposed that the scene images consist of the objects and their relations, and proposed the scene constitution model based on image regions using the graphical model. They measured a similarity between graphs so that they accomplished semantic image classification and retrieval. Cheng and Wang [3] proposed the contextual Bayesian network (CBN) model, where the hybrid streams of object occurrence and spatial relations between objects are piped into the CBN-based inference engine. Vogel et al. [11] proposed a concept occurrence vector (COV) for semantic image description, where the image blocks were classified into object concepts and the whole image has been expressed by frequency of occurrence of the concepts. The related works described above have their own inherent characteristics in semantic image description, but all have same weakness, i.e. the semantic image description that consists of objects and their relations was considered in vector space, so similarity measure methods also used the distance in vector space. The representation of semantic image information is just the description of knowledge. Li et al. [5] described semantic image features using a linguistic variables defined as 5-tuple $<x, T(x), U, G, M>$. Here x is a name of linguistic variables, $T(x)$ is a set of value of linguistic variables (natural language predicate), U is a domain of definition, G is a grammar rule of the extended predicates that appear in $T(x)$, and M is a semantic rule of fuzzy degree that gets the variable value. Expressing the semantic information with the linguistic variables, it can conveniently describe the image contents. However they used description of the texture image features, they did not extend to semantic features else. Zhang et al. [13] made a semantic image understanding model using the generalized set. They defined the generalized set considered of a set of elements such as numeral value, vector, sentence, and so on, a set of rules that generalized set is formed by them, and a set of operational rules. And they discussed the extraction and understanding of semantic information of image based on the generalized set. Generalized set fuses numeral value, vector and sentence as one, so that it has more strong description ability. Employing the generalized set for semantic image representation with generalized space instead of the original feature space, they could effectively carry out the semantic image understanding and analysis. For semantic image classification, Bayesian classifier and support vector machine (SVM) classifier are widely used.

Aksoy et al. [1] proposed the three-level visual grammar model and classified the pixels. For image classification, Papadopoulos et al. [8] employed the SVMs using global and local features of image. Qi et al. [9] considered the inner-class difference of an image class and proposed the method to construct the prototype set of the spatial contextual models by employing kernel methods. Then they carried out the image classification using the distance measure between image models. Vogel et al. [11] implemented the image classification using SVM classifier based on COV distance. Since the traditional SVM classifier could be solved a binary classification problem, researchers used a multi-classifier that consists of several SVMs: one-against-rest and one-against-one. For the multiclass (n classes) classification problem, one-against-rest strategy [12] trains the n SVMs, then the separation plane of i^{th} SVM classify the data into class i or non-class i. However, in this case a data has been included in several classes, or there is a case that is not included in any classes. To solve this problem, researchers used the fuzzy classification strategy [6]. In this paper, we employ the same SVM classifier as the work of Vogel et al. [11]. So the classification model did not change much with the related works, but it is not the same as the distance measurement method. In this paper, we employ one-against-rest strategy where the n SVMs (n = 6: the number of image classes that should be classified) are

trained. Further, we carry out the multi-class classification based on semantic distance between images.

Based on above discussion, we consider that image knowledge could be expressed with different data types (for example numeral value, vector, matrix, character, and sentence etc.), so the semantic image classification has to employ the generalized set [13]. This is an important way to reduce the semantic gap [14]. We have been inspired by the researches of Zhang and Zhu. In order to inherit and evolve their researches, we propose the representation of image knowledge and its actual application in image classification.

The contributions of this paper are as follows:

- We propose the image description model which includes various types of data (vector, sentence, and symbol) and define the semantic similarity measure between images. This image description model is similar to generalized set based scheme proposed by Zhang et al. [13]. It is more semantic representation scheme of image knowledge than other vector or graph based model. Under the image description model based on generalized set, we propose the semantic similarity measurement which is implemented in various spaces and is fused into one type. It is just different point from other generalized set based methods and is the one contribution.
- We propose the semantic image classification method by employing multi SVM classifiers based on semantic distance. In previous works, SVM classifier was widely used to classify image semantics and confirmed its good performance. We employ one-against-rest strategy of SVMs to solve multi classification problem. Here we use RBF kernel constructed by using the semantic distance between images instead of the vector distance. It is the other one contribution. Finally, through experiments, we illustrate the validity of proposed method.

2. SEMANTIC IMAGE DESCRIPTION BASED ON GENERALIZED SET

The image description model can be formally defined as a 5-tuple

$$I = <ic, V, C, S, R> \quad (1)$$

where, *ic* is a semantic label of image, *V* is a set of visual features of image, *C* is a set of semantic concepts of objects included in an image, *S* is a set of contextual relations between objects, and *R* is a rule that maps a content of image into semantic label.

First of all, we discuss the objects included in an image. We assume that the objects included in a specific domain image are finite. In the image description model based on generalized set, *C* is a set of semantic concepts of objects, i.e. the finite set of words. We describe the outside natural scene images with fourteen semantic objects: *C* = {sky, water, grass, plant, rock, sand, mountain, snow, ground, building, person, vehicle, things, animal}. With this fourteen object concepts, 99.3% of considered domain (coast, forest, highway, mountain, field and street) images could be labeled.

The context between objects included in an image can be grouped into three categories [4]: semantic context, spatial context, and scale context. In the image description model based on generalized set, *S* is the set of relations between objects, i.e. the finite set of characters. Semantic context and scale context are related with the objects appeared in an image and their sizes. The objects appeared in an image are included in the set of object concepts. Spatial context between objects can be divided into three types: directional relations, distance relations and topological relations. In this paper we consider the characteristics of the natural scene images, i.e. left direction and right one don't affect the image classification, and we deal with the relationship between the adjoined objects.

The object concepts in image are obtained from the visual features. In addition, we couldn't consider the contextual relations between objects without the concepts of objects. So the visual features of image are very important. Previous content-based image processing (for example content-based image classification and retrieval) was usually accomplished using the global visual features of image. The visual features of image (color, texture, and shape features et al.) are expressed with the vectors generally, i.e. $V = <fcolor, ftexture, fshape>$.

Therefore, the model that includes various types of data (vector, word, character) has the ability to describe the image semantics perfectly, and it has been corresponded to human perception. For semantic image understanding and analysis it is very important how to make the mapping rule *R* in the model. The mapping rule is related to semantic similarity measure between images. Since the image description model is based on generalized set, the similarity between images are got from the similarity measure in set theory: $sim(A, B)=N(A\cap B)/N(A\cup B)$, where *A*, *B* are two sets, $N(\cdot)$ is the number of set's elements.

We have made the model of the image semantics using the generalized set, therefore the similarity between image knowledge is measured by computing the similarity between the models of generalized set. For images I_1, $I_2 \in F$ (V, C, S), the similarity is defined following:

$$sim(I_1, I_2) = <sim_V(I_1, I_2), sim_C(I_1, I_2), sim_S(I_1, I_2)> \quad (2)$$

There are tree difference sets, so the similarity between models has three parts, too. We discuss the similarity measure of every set type respectively.

Euclidian distance has been widely employed in vector space. For a visual similarity measure between images, we employ the Euclidian distance:

$$sim_V(I_1, I_2) = 1 / | V_1 - V_2| \quad (3)$$

Here we have considered that the similarity and the distance are inversely proportional.

For a concept set, we employ the following similarity measure in set theory:

$$sim_C(I_1, I_2)= N(C_1\cap C_2)/N(C_1\cup C_2) \quad (4)$$

Here C_1 and C_2 are the sets of semantic concepts included in images I_1 and I_2. Considering the important of contents, we change it as follows:

$$sim_C(I_1, I_2) = \frac{\sum_{c_i \in C_1 \cap C_2} \min(w_{1i}, w_{2i})}{\sum_{c_j \in C_1 \cup C_2} \max(w_{1j}, w_{2j})} \quad (5)$$

Here w_{1i} and w_{2i} are the weights of the concept c_i which represent the importance of objects c_i in images I_1 and I_2. They are calculated as a percentage of the area occupied by the objects in an image.

The similarity between two character sets could be measured as follows:

$$sim_S(I_1, I_2) = N(S_1 \cap S_2)/N(S_1 \cup S_2) \quad (6)$$

In this case, we don't consider the important degree of characters. This is because here only the presence or absence of contextual relationships between objects in the images is considered.

Outputs of these three similarity measures have numerical format, i.e. their ranges are normalized to [0,1] respectively. If two images are more similar, then the similarity value is closer to 1; in contrast, if two images are completely dissimilar, then its one is 0.

These three similarity measures are fused into one form. In this paper, we use a simple weighted linear combination to calculate the entire similarity between two images.

$$sim(I_1,I_2) = a_V \cdot sim_V(I_1,I_2) + a_C \cdot sim_C(I_1,I_2) + a_S \cdot sim_S(I_1,I_2) \quad (7)$$

Here a_V, a_C and a_S are the corresponding weights of sim_V, sim_C and sim_S respectively, and their sum is 1. In this paper, these weight values are simply 1/3.

3. SEMANTIC IMAGE CLASSIFICATION

Semantic image classification has included the definition of semantic similarity measure and the design of classifier. An outside natural scene image is composed of several objects and their features are affected by weather, season etc. Therefore its classification by computer is difficult. For instance, even the same trees or plants have different color features in different seasons. The sea, river and lake could not be distinguished correctly. However, human being correctly distinguishes and classifies them. Because human perception depends on domain knowledge which includes the presences of objects and contextual information between objects.

In this paper we carry out the semantic classification of the outside natural scene images using the image description model based on generalized set. We use fourteen objects concepts to describe six natural scene classes. For image classification, we employ the SVM classifier with the semantic image similarity between images.

Let $\{ITR_i\}$ is the set of training images, *ITE* is the test image. SVM is one of the pattern classification methods based on the statistical training theory. It solves the non-linear and high dimensional problems in pattern classification. In the case of binary linear classification, the separation plane separates the training data into two classes. Here the separation plane gives the maximum margin between the plane and the support vectors of each class. However, in practical problems, there are many situations in which linear classifiers are not classified. For non-linear case, the kernel function is employed. Using the kernel function $K:(R^n:R^n) \rightarrow R, K(x_i, x_j)$ is equal to inner product of x_i, x_j of the high dimensional space. We employ the Radial basis function (RBF) as kernel function of SVM:

$$K\left(x_i, x_j\right) = \exp\left\{-\frac{\left|x_i - x_j\right|^2}{\sigma^2}\right\} \quad (8)$$

Here distance is the Euclidian distance. We use the semantic distance between images instead of the vector distance, then we have the RBF kernel as follows:

$$K\left(I_i, I_j\right) = \exp\left\{-\frac{d_{SIMANTIC}\left(I_i, I_j\right)}{\sigma^2}\right\} \quad (9)$$

Here the semantic distance, $d_{SIMANTIC}$, between two images is obtained from semantic similarity mentioned above:

$$d_{SIMANTIC} = 1 - \text{sim}(I_i - I_j) \quad (10)$$

This distance value also belongs to the [0, 1] range.

Since the traditional SVM is proposed for binary classification problem, we use the multiclass SVMs that consist of several SVMs. We employ one-against-rest strategy, where six SVMs are trained, and then we execute the multi-class classification based on semantic distance between images.

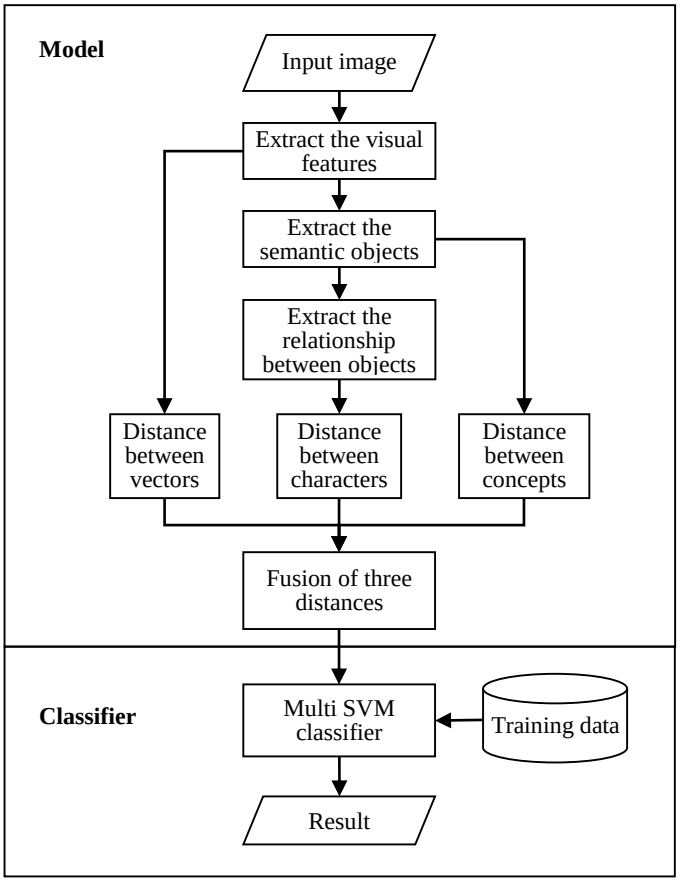

Fig.1. Flow diagram of the semantic image classification

The Fig.1 shows the flow diagram of the semantic image classification. Semantic image classification scheme composed of two parts: semantic image description model presented in section 2 and the image classification using the multi SVMs based on above model. In classifier part of Fig.1, the training database is also constructed by image description model based on generalized set.

Proposed image classification method is based on semantic similarity measure between images, where the similarity measure reflects the semantic objects information and their relationships. Different from the visual feature based low-level image classification (or content based image classification); in this

paper, we use not only the visual features but also the semantic objects information and their relationships.

4. EXPERIMENTAL RESULT

In order to verify the effectiveness of image description and classification methods which are based on generalized set, we carried out the related experiment and compared the results with other advanced ways. On the other hand, we compared with the image classification method based on the global visual features. The semantic image description model based on concept occurrence vector (COV) proposed by Vogel (2007) is one of the state-of-the-art baselines. They employed SVMs for image classification and its performance is relatively well. Another state-of-the-art baseline is the image description model based on contextual Bayesian network (CBN) proposed by Cheng et al. (2010). We used HSV color histogram and wavelet texture features as global visual features of image.

Table.1. Confusion matrix of the image classification based on the global visual features

	coa	for	hig	mou	fie	str
coast	**61.76**	2.21	11.03	14.71	8.82	1.47
forest	0	**90.22**	0.89	0	8.00	0.89
highway	13.91	4.64	**65.56**	9.27	2.65	3.97
mountain	8.39	8.76	9.85	**43.07**	21.17	8.76
field	10.61	14.01	7.58	8.71	**56.82**	2.27
street	0.53	4.71	0.52	1.05	1.57	**91.62**

Table.2. Confusion matrix of the image classification based on COV [11]

	coa	for	hig	mou	fie	str
coast	**98.53**	0.74	0	0.73	0	0
forest	0.45	**96.89**	0.44	0.89	1.33	0
highway	0	0	**92.72**	0	1.32	5.96
mountain	0.37	1.09	0.73	**92.70**	5.11	0
field	4.55	6.44	8.71	9.09	**70.83**	0.38
street	0	0	5.76	0	0	**94.24**

Table.3. Confusion matrix of the image classification based on CBN [13]

	coa	for	hig	mou	fie	str
coast	**99.26**	0.74	0	0	0	0
forest	0.45	**96.89**	0.44	0.89	1.33	0
highway	0	0	**95.37**	0	0.66	3.97
mountain	0.37	0.73	0.36	**93.80**	4.74	0
field	4.92	2.27	7.58	10.98	**73.11**	1.14
street	0.52	0	2.62	0	1.05	**95.81**

Table.4. Confusion matrix of the image classification based on proposed method in this paper

	coa	for	hig	mou	fie	str
coast	**99.27**	0.73	0	0	0	0
forest	0.45	**98.22**	0.44	0	0.89	0
highway	0	0	**91.39**	0	0.66	7.95
mountain	1.09	1.10	0.73	**96.35**	0.73	0
field	5.30	3.03	7.58	10.98	**71.97**	1.14
street	0	0.52	1.57	0	0	**97.91**

We used the image data set spatial_envelope_256x256_static_8outdoor categorie provided by LabelMe (http://labelme.csail.mit.edu/). The image data set contains 1841 images of 256 × 256 pixels resolution and is classified into 6 classes manually: coast (236 images), forest (325 images), highway (251 images), mountain (374 images), field (364 images), and street (291 images). In each image class, we selected 100 images (there are all 600 images) as training data, and the rest of images regarded as test data (there are all 1241 images). For fair comparison, all the experiments are executed by a 10 cross-validation process on different training sets and test sets and the average values of them are reported. The classification performance is evaluated with the classification accuracy and the confusion matrix. The Fig.2 showed the comparison of classification accuracy. The Table.1 expressed the confusion matrix of the image classification based on the global visual features, and Table.2, Table.3 and Table.4 expressed the confusion matrix of semantic image classification based on the COV, CBN, and proposed method in this paper, respectively. Consequently, the classification accuracy of the proposed method is better than the other ways: In the six image classes, the classification accuracy of the proposed method is higher than ones based on global visual features. This is because of existence of the semantic gap between visual features and human perception. Comparing with COV-based method, the average classification accuracy of the proposed method was improved to 1.53%. This is because Vogel used the vector space distance (SSD) and we used the semantic distance between image models. Comparing with CBN-based method, the classification capability of the proposed method is almost similar with CBN-based method and average classification accuracy is higher 0.15%. For semantic image description and classification, Cheng et al. [3] considered the spatial relations between objects in an image. Through the experiment and comparison of classification performance, we have verified the suitability and effectiveness of the semantic image description model and the semantic similarity measure method based on generalized set proposed in this paper.

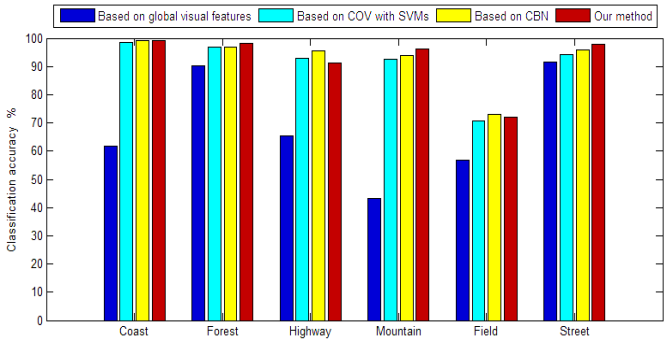

Fig.2. Comparison of image classification accuracy

5. CONCLUSION

In order to describe the image semantics, in this paper proposed an image description model that uses not only the visual features of image but also the information of objects included in image and the relationship between the objects. We have used the concept of a generalized set which integrate various information. Based on generalized set, we have proposed a similarity measure method between images. We have also performed image classification by employing SVM classifier based on semantic distance between images. Proposed image description model represented the image contents completely, and is similar to human perception. The experimental result demonstrated that the image classification method based on semantic similarity measure improved the classification accuracy. However, extracting semantic objects from images is a difficult problem. For natural image, the segmentation accuracy is not sufficient, so researchers are investigating ways to recognize the objects without image segmentation. And in this paper, we did not solve the fusion of three different similarities well. We used a simple weighted linear combination to calculate the entire similarity between two images. This is also one of the important problems in the semantic image processing. In future work, we will study the fusion of different information types.

REFERENCES

[1] S. Aksoy, et al., "Learning Bayesian Classifiers for Scene Classification With a Visual Grammar", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 43, No. 3, pp. 581-589, 2005.

[2] M.R. Boutell, J. Luo and C.M. Brown, "Scene Parsing using Region-Based Generative Models", *IEEE Transactions on Multimedia*, Vol. 9, No. 1, pp. 136-146, 2007.

[3] H. Cheng and R. Wang, "Semantic Modeling of Natural Scenes based on Contextual Bayesian Networks", *Pattern Recognition*, Vol. 43, No. 12, pp. 4042-4054, 2010.

[4] C. Galleguillos and S. Belongie, "Context Based Object Categorization: A Critical Survey", *Computer Vision and Image Understanding*, Vol. 114, No. 6, pp. 712-722, 2010.

[5] Q. Li, H. Hu and Z. Shi, "Semantic Feature Extraction using Genetic Programming in Image Retrieval", *Proceedings of 17th International Conference on Pattern Recognition*, pp. 302-309, 2004.

[6] G. Madzarov, D. Gjorgjevskij and I. Chorbev, "A Multi-Class SVM Classifier Utilizing Decision Tree", *Informatica*, Vol. 33, No. 2, pp. 233-241, 2009.

[7] I. Nwogu, V. Govindaraju and C. Brown, "Syntactic Image Parsing using Ontology and Semantic Descriptions", *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 41-48, 2010.

[8] G.T. Papadopoulos, V. Mezaris, I. Kompatsiaris and M.G. Strintzis, "Combining Content and Context Information for Semantic Image Analysis and Classification", *Proceedings of 15th European International Conference on Signal Processing*, pp. 708-712, 2007.

[9] G.J. Qi, X.S. Hua, Y. Rui, J. Tang and H.J. Zhang, "Image Classification with Kernelized Spatial-Context", *IEEE Transactions on Multimedia*, Vol. 12, No. 4, pp. 278-287, 2010.

[10] K.E. Sanders, B.P. Kettler and J.A. Hendler, "The Case for Graph-Structured Representations", *Proceedings of International Conference on Case-Based Reasoning*, pp. 245-254, 1997.

[11] J. Vogel, "Semantic Modeling of Natural Scenes for Content-based Image Retrieval", *International Journal of Computer Vision*, Vol. 72, No. 2, pp. 133-157, 2007.

[12] J. Yang, K. Yu, Y. Gong and T. Huang, "Linear Spatial Pyramid Matching using Sparse Coding for Image Classification", *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 163-169, 2009.

[13] Y. Zhang, M. Yao and Z. Yuan, "Research on Methodology of Image Semantic Understanding based on Generalized Computing", *Proceedings of IEEE International Conference on Pattern Recognition and Computer Vision*, pp. 113-117, 2007.

[14] R. Zhu, M. Yao and Y. Liu, "Image Classification Approach Based on Manifold Learning in Web Image Mining", *Proceedings of International Conference on Advanced Data Mining and Applications*, pp. 780-787, 2009.