Scene Classification via Hypergraph-Based Semantic Attributes Subnetworks Identification

Sun-Wook Choi, Chong Ho Lee, and In Kyu Park

Department of Information and Communication Engineering
Inha University, Incheon 402-751, Korea
swchoi@inhaian.net, {chlee,pik}@inha.ac.kr

Abstract. Scene classification is an important issue in computer vision area. However, it is still a challenging problem due to the variability, ambiguity, and scale change that exist commonly in images. In this paper, we propose a novel hypergraph-based modeling that considers the higher-order relationship of semantic attributes in a scene and apply it to scene classification. By searching subnetworks on a hypergraph, we extract the interaction subnetworks of the semantic attributes that are optimized for classifying individual scene categories. In addition, we propose a method to aggregate the expression values of the member semantic attributes which belongs to the explored subnetworks using the transformation method via likelihood ratio based estimation. Intensive experiment shows that the discrimination power of the feature vector generated by the proposed method is better than the existing methods. Consequently, it is shown that the proposed method outperforms the conventional methods in the scene classification task.

Keywords: Scene classification, Semantic attribute, Hypergraph, SVM.

1 Introduction

Scene understanding still remains a challenging problem in computer vision field. Among particular topics in scene understanding, image classification including scene and object classification has become one of the major issues. Over the past decade, numerous techniques have been proposed to classify scene images into appropriate categories.

Most of the existing high-level image classification techniques are performed in the transformed domain from the original image. Popular approaches employ the statistics of local feature such as histogram of textons [15] and bag-of-words (BoW) [6,24]. In the BoW model, local features obtained from an image are first mapped to a set of predefined visual words, which is done by vector quantization of the feature descriptors using a clustering technique such as K-means. The image is then represented by a histogram of visual words occurrence. The BoW model has demonstrated remarkable performance in challenging image classification tasks when it is combined with the well known classification techniques such as the support vector machine (SVM) [9,12,28].

D. Fleet et al. (Eds.): ECCV 2014, Part VII, LNCS 8695, pp. 361–376, 2014.
© Springer International Publishing Switzerland 2014

However, the conventional methods using low-level feature have a few limitations. First, although the visual words are more informative than individual image pixels, they still lack explicit semantic meanings.

Second, the visual words are occasionally polysemous, so it is possible to have different semantic meanings even though we have the identical visual word [25].

To overcome the limitations, several techniques have been proposed so far. Bosch *et al.* utilizes the intermediate-level representation to shows the improved classification performance [3]. However, the intermediate-level image representation technique is yet not free from the problem of visual words' ambiguity, *i.e. polysemy* and *synonymy* [25].

Recently, in order to utilize the higher-level contextual information for scene classification, semantic attribute is investigated actively. This technique alleviates the effect of the *polysemy* and *synonymy* problems if it is combined with the contextual information correctly [19,22,25]. Note that the semantic attributes in a scene are intuitive. Usually, they represent individual objects in the scene as well as the particular parts of them. In general, these semantic attributes show higher-order relationship between each other, which can be usually observed in real scenes. For example, a 'street' scene can be thought of as a combination of a 'building', 'road', and 'car'. However, the existing techniques using high-level semantic attributes do not exploit the higher-order relationship adequately but treat the semantic attributes independently. Furthermore, it is noticeable that some semantic attributes co-occur frequently while some other semantic attributes rarely appear together in a certain scene. We can represent the co-occurrence as the relation of semantic attributes. The interaction (relations) of semantic attributes provides strong contextual information about a scene. Based on this idea, we attempt to exploit the higher-order interaction of semantic attributes for the scene classification problem.

Generally, a graph-based modeling technique is can be considered to deal with the interaction between attributes. However, typical graph-based models use formulation that involve only single pairwise interactions and are not sufficient to model the higher-order interaction. To overcome this limitation, we can consider a hypergraph-based technique to model the higher-order interaction.

A hypergraph is a generalization of the conventional graph structure in which a set of nodes is defined as hyperedge [26]. Unlike the conventional graph model, a hypergraph contains the summarized local grouping information represented by hyperedges. In the hypergraph model, it is possible to construct various combinations of hyperedges using different sets of attributes. These hyperedges co-exist in a hypergraph and provide complementary information for the target data. In this context, the hyperedges are regarded as subnetworks of attributes. This property is beneficial to model the co-occurrence patterns of semantic attributes in scene images.

In this paper, we propose a hypergraph-based scene modeling and learning method for scene classification. By employing the hypergraph learning, we can search important subnetworks on the efficiently from the interaction network of the semantic attributes. Overall process of proposed method for scene

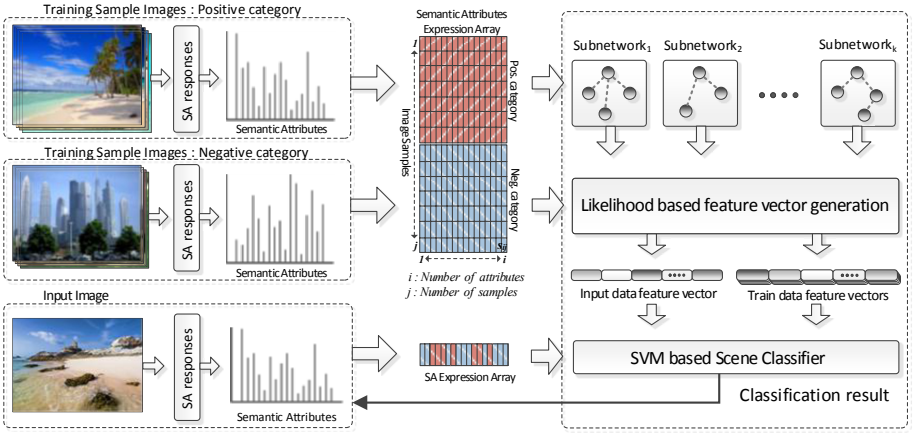

Fig. 1. Overall process of proposed method. Given training images and an input image, expression arrays of semantic attributes are calculated from responses of semantic attributes for each image. Then, for each scene category, category-specific semantic attribute (CSSA) subnetworks are obtained by proposed hypergraph based method. New feature vectors are generated by aggregation method from explored subnetworks and then input image is classified based on the newly generated feature vector with a SVM based scene classifier.

classification is depicted in Fig. 1. The main contributions of this paper are summarized as follows.

1. In order to take account the higher-order interaction of semantic attributes in scene classification, we propose a novel scene classification method based on the hypergraph model and efficient learning technique to create the category-specific hypergraph suitable for scene classification problem.
2. We propose another novel method to generate aggregated feature vector from the hypergraph suitable for discriminative model. The newly generated feature vector not only reduces the dimension of the feature space, but also alleviates the measurement noise of semantic attributes expression and therefore. This enables us to obtain a robust feature vector.

2 Hypergraph-Based Scene Modeling

In order to model a particular scene category using a hypergraph, we use the co-occurrence pattern of semantic attributes obtained from either the training sample images in the scene category or the text-based annotation of scene images. For example, when we build the hypergraph of the 'coast' category, we can get hyperedges from the training sample images I included in the category. Assuming that the number of training sample images is six, the set of semantic attributes S is generated for the 'coast' category as follows.

$$S = \{s_1(\text{`}sky\text{'}),\ s_2(\text{`}water\text{'}),\ s_3(\text{`}sand\text{'}),\ s_4(\text{`}boat\text{'}),\ s_5(\text{`}tree\text{'}),\ s_6(\text{`}rock\text{'})\}. \quad (1)$$

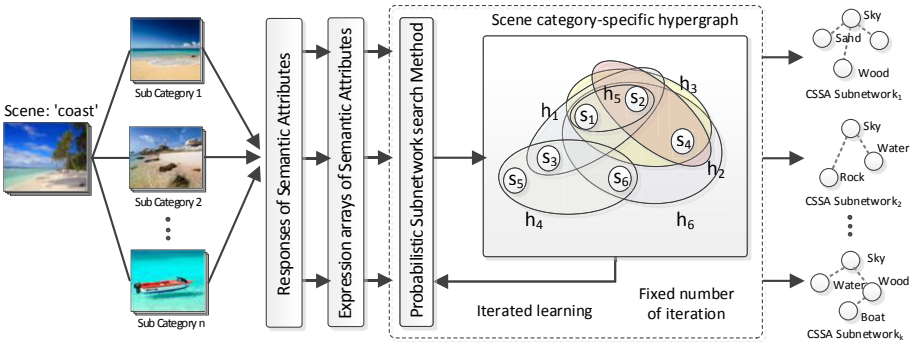

Fig. 2. A hypergraph model based on semantic attributes for category-specific scene modeling. The hypergraph model is optimized by a population-based evolutionary learning method to obtain CSSA subnetworks for scene classification.

We consider these semantic attributes S as the set of nodes of a graph. The Hyperedges E can be obtained from the image set $\boldsymbol{I}$ as follows.

$$E = \{e_1 = \{s_1, s_2, s_3\}, e_2 = \{s_2, s_4\}, e_3 = \{s_1, s_2, s_4\}, \\ e_4 = \{s_3, s_5, s_6\}, e_5 = \{s_1, s_2\}, e_6 = \{s_1, s_2, s_4, s_6\}\}. \tag{2}$$

The hypergraph $H = (S, E)$ consists of the set of nodes S and the set of hyperedges E. As in the example above, we can build a hypergraph for the certain scene category by combining the hyperedges obtained from its training sample images. Due to this characteristic, the hypergraph can be represented by the population which consists of the hyperedge. Here, each hyperedge is considered as each individual of the population.

As shown in Eq. (2), a hypergraph can model edges including an arbitrary number of attributes. Based on this property of a hypergraph, we can model each scene category by means of a hypergraph. However, even though the hypergraph model represents a certain scene category very well, it is yet still insufficient for the scene classification task. This is because the relative distribution between the desired category to be classified (denoted as 'positive category' in Fig. 1) and other categories (denoted as 'negative category' in Fig. 1) is not considered when building hyperedges of a hypergraph. Therefore, a generation of appropriate hypergraph for a scene classification task means that it is a searching process of suitable hyperedges to represent the characteristics of the target category efficiently as well as to consider the discrimination capability from other categories.

For this reason, a learning process is required to refine the initial hypergraph. In our approach, we employ the population-based learning model based on [29]. Here, the hypergraph H is re-defined as $H = (S, E, W)$ where S,E, and W are a set of vertices (semantic attributes), hyperedges, and weights of each hyperedge, respectively. That is, the re-defined hypergraph adds weight terms to the original hypergraph. Each vertex corresponds to a particular semantic attribute

while each hyperedge represents the relational combination of more than two vertices with its own weight. The number of vertices in a hyperedge is called the cardinality or the order of a hyperedges, in which k-hyperedge denotes a hyperedge with k vertices.

Since the hypergraph with the weight term can be regarded as a probabilistic associative memory model to store segments of a given data set $D = \{x^{(n)}\}_{n=1}^{N}$ *i.e.* $\mathbf{x} = \{x_1, x_2, ..., x_m\}$ as in [29], a hypergraph can retrieve a data sample after the learning process. When $I(x^{(n)}, E_i)$ denotes a function which yields the combination or concatenation of elements of E_i, then the energy of a hypergrpah is defined as follows.

$$\begin{aligned} &\varepsilon(\mathbf{x}^{(n)}; W) = -\sum_{i=1}^{|E|} w_i^{(k)} I(\mathbf{x}^{(n)}, E_i) \\ &\text{where} \quad I(\mathbf{x}^{(n)}, E_i) = x_{i1}^{(n)} x_{i2}^{(n)} ... x_{ik}^{(n)}. \end{aligned} \tag{3}$$

In Eq. (3), $w_i^{(k)}$ is the weight of the i-th hyperedges E_i with k-order, $\mathbf{x}^{(n)}$ means the n-th stored pattern of data, and E_i is $\{x_{i1}, x_{i2}, ..., x_{ik}\}$. Then, the probability of the data generated by a hypergraph $P(D|W)$ is given as a Gibbs distribution as follows.

$$P(D|W) = \prod_{n=1}^{N} P(\mathbf{x}^{(n)}|W), \tag{4}$$

$$P(\mathbf{x}^{(n)}|W) = \frac{1}{Z(W)} exp(-\varepsilon(\mathbf{x}^{(n)}; W)), \tag{5}$$

where $Z(W)$ is a partition function, which is formulated as follows.

$$Z(W) = \sum_{\mathbf{X}^{(m)} \subset D} \exp\left\{ \sum_{i=1}^{|E|} w_i^{(k)} I(\mathbf{x}^{(m)}, E_i) \right\}. \tag{6}$$

That is, a hypergraph is represented with a probability distribution of the joint variables with weights as parameters when we consider attributes in data as random variables.

Considering that learning of a hypergraph is to select hyperedges with a higher weight, it can be formulated as the process for maximizing log-likelihood. Learning from data is regarded as maximizing probability of weight parameter of a hypergraph for given data D. In this context, the probability of a weight set of hyperedges $P(W|D)$ is defined as follows.

$$P(W|D) = \frac{P(D|W)P(W)}{P(D)}. \tag{7}$$

According to Eq. (5) and Eq. (7), the likelihood is defined as

$$\prod_{n=1}^{N} P(\mathbf{x}^{(n)}|W)P(W) = \left(\frac{P(W)}{Z(W)}\right)^{N} \exp\left\{ -\sum_{n=1}^{N} \varepsilon(\mathbf{x}^{(n)}; W) \right\}. \tag{8}$$

Ignoring $P(W)$, maximizing the argument of exponential function is equivalent to obtaining maximum log likelihood as follows.

$$\begin{aligned} &\arg\max_{W}\left[\log\left\{\prod_{n=1}^{N} P(\mathbf{x}^{(n)}|W)\right\}\right] \\ &= \arg\max_{W}\{\sum_{n=1}^{N}\sum_{i=1}^{|E|} w_i^{(k)} I(\mathbf{x}^{(n)}, E_i) - N\log Z(W)\}. \end{aligned} \tag{9}$$

More detail derivation of the log-likelihood are shown in [29]. A likelihood function can be maximized by exploring different hyperedge compositions which can reveal the distribution of given data better. Now, the problem is converted to finding appropriate combination of optimal hyper edges (or feature subsets). In other words, this is equivalent to exploring a suitable hyperedge-based population for a scene classification.

In general, exploring optimal feature subsets from a high-dimensional feature space is an NP-complete problem since it is impractical to explore the entire feature space. In this paper, we propose a sub-optimal hypergraph generation method to maximize the discrimination power from the perspective of Kleinberg's stochastic discrimination (SD) [13]. Based on the central limit theorem, the SD theory proves it theoretically and demonstrated experimentally that it is possible to generate a strong classifier by producing and combining many weak classifiers based on randomly sampled feature subsets. However, in order to approximate the actual distribution of data, it requires repeated random sampling process for an entire feature space as many as possible. For an efficient search, we use a heuristic search based on a population-based evolutionary computation technique as illustrated in Fig. 2. The details will be explained in the following section.

3 Learning of a Hypergraph for Scene Classification

To learn the hyper graph, we use a population-based evolutionary learning method. In this method, the variation, evaluation, and selection are performed iteratively. In the proposed learning method, a hyperedge is weighted by (i) the amount of higher-order dependency and (ii) the discrimination power between different categories. As the population changes in the learning procedure, the hypergraph structure evolves by removing hyperdges with relatively low weight and by replacing them with new hyperedges with relatively high weight. Filtering is subsequently performed for each hyperedge to have optimal elements.

In general, features on search space have the equivalent selection probability under the uniform distribution. However, it is inefficient because features used in classification do not have equivalent discriminative power. Furthermore, in the scene classification problem, it has the characteristic that the occurrence probability of each feature in the entire feature space is very sparse. Therefore, we need to adjust selection probability of each feature based on the importance of each feature.

Algorithm 1. Two-step sub-optimal hypergraph learning algorithm (Step 1)

Input: Expression array $\boldsymbol{A}$ of semantic attributes
Output: Candidate group of attribute subsets $\boldsymbol{g}$

1: $\omega(s^i) \leftarrow$ Weight calculation of each attribute s^i by Student's t-test.
2: $\boldsymbol{g} \leftarrow$ Generate a fixed number of m semantic attribute subsets randomly from original semantic attribute space based on the weight $\omega(s^i)$.
3: **repeat**
4: (a) $f(g^k) \leftarrow$ Calculate a fitness value of the subset g^k.
5: (b) Sort semantic attribute subsets g^k in descending order based on fitness values of $f(g^k)$ given by Eq. (10).
6: (c) Remove the bottom 30%, and then replace with newly created subsets.
7: **until** (Predetermined number of learning.)

Algorithm 2. Two-step sub-optimal hypergraph learning algorithm (Step 2)

Input: Candidate group of semantic attribute subsets $\boldsymbol{g}$
Output: Sub-optimized subsets $\hat{\boldsymbol{g}}$

1: **for** k = 1 to $|\boldsymbol{g}|$ **do** : for each subset
2: Sort member attributes s^i in descending order based on the abs. t-test score of s^i.
3: Add the 1st ranked s^i to empty set $\hat{g}^k$, then calculate the discriminative power d^k.
4: **repeat**
5: (a) $\hat{g}^k = \hat{g}^k \cup \{s^i\}$: Add the next ranked s^i to subset $\hat{g}^k$.
6: (b) Evaluate the discriminative power d^k of $\hat{g}^k$.
7: **until** (The new d^k is less than the previous d^k.)
8: **end for**

To solve this problem, we propose a probabilistic subnetworks search method that can sample the feature subsets efficiently based on the importance of each feature. In order to generate subnetworks of semantic attributes for scene classification using the probabilistic subnetworks search method, first, we need to measure the importance of each feature. This can be measured by the degree of association with the target category. In case of continuous data with a small number of samples, we can consider to use Student's t-test for two-class problem [8]. Using the t-test, we can get P-value which is the rejection probability of the null hypothesis $H_0 : \mu_p = \mu_n$ assuming that the mean of each population belonging to the positive category and negative category is equal. P-value has high rejection probability of the null hypothesis when its value is close to 0. This means that the distribution pattern of each class is very different. In other words, the discriminative power of each feature is strong. In order to accommodate this to the selection probability of semantic attribute s_i, we use the value of $1 - p(s_i)$ where $p(s_i)$ is a P-value of s_i. Strictly speaking, it is not the probability score,

but it has the property that the selection capability becomes high when it close to 1 and vice versa.

The subnetworks search process based on the selection probability of semantic attributes is shown in Algorithm 1. At the beginning of the search, we create the predefined number of subsets (m=1000), then obtain the fitness values of each subset. Among various techniques, we use Hotelling's T^2-test [21] used in multivariate test to obtain a robust fitness value with fast speed. Note that Hotelling's T^2-test is a generalization of Student's t-test that is used in multivariate hypothesis testing. It is suitable for assessing the statistically higher-order relationship of semantic attributes composing the subset, since it can consider the correlation and interdependence between the component of subset.

Hotelling's T^2-test score for generated subsets is calculated as follows.

$$T^2 = \frac{n_p n_n}{n_p + n_n}(\overline{\mathbf{A}}^{\mathbf{p}} - \overline{\mathbf{A}}^{\mathbf{n}})\mathbf{S}^{-\mathbf{1}}(\overline{\mathbf{A}}^{\mathbf{p}} - \overline{\mathbf{A}}^{\mathbf{n}})^{\top}, \tag{10}$$

where n_c is the number of samples belongs to each category c, $\overline{A}^c$ is the mean expression value of semantic attributes s_j^p and s_j^n belongs to each category. $\mathbf{S}$ is the pooled variance-covariance matrix of semantic attributes.

In the probabilistic subnetworks search method, subsets are generated through the search process as shown in Algorithm 1 using a fitness function based on the Hotelling's T^2-test. Then, filtering is performed for each subset to have sub-optimal member attributes that maximizing discriminative power through the incremental learning as shown in Algorithm 2. The sub-optimal subsets obtained from the search process are regarded as subnetworks because its member attributes are likely to interact each other. Finally, these subnetworks are hyperedges which build the learned hypergraph.

4 Feature Vector Generation Based on Likelihood Ratio

In order to use the learned hypergraph for scene classification, we employ a discriminative model. When using a discriminative model, the parameter learning is relatively simpler than a generative model. In addition, it has an advantage of an ease to utilize well-known classification methods such as support vector machine (SVM) that are known to show relatively superior classification capability in many fields.

For using the discriminative model, we need generation of feature vectors from a learned hypergrpah. Each hyperedge constituting the hypergraph includes multiple semantic attributes as shown in Fig. 2. Thus, it is impossible to apply to the classification model directly. For this, we propose a method based on the likelihood ratio to aggregate the expression values of each member attribute to make up the hyperedges from the original expression data.

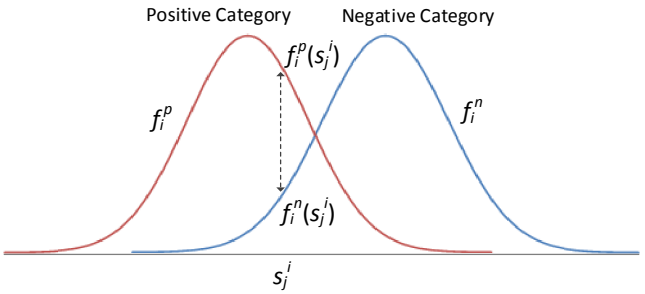

Fig. 3. Likelihood estimation for each semantic attribute

Given a expression vector $\mathbf{s}_j = (s_j^1, s_j^2, ..., s_j^n)$ which contains the expression levels of the member semantic attributes, we estimate the aggregate value of g_j^k(k-th subnetworks of sample j) as follows.

$$aggregated\ value\ of\ g_j^k = \frac{1}{|g_j^k|} \sum_{i=1}^{n} \lambda_i(s_j^i), \tag{11}$$

where $\lambda_i(s_j^i)$ is the likelihood ratio between positive and negative categories for the semantic attributes. The likelihood ratio $\lambda_i(s_j^i)$ is given by

$$\lambda_i(s_j^i) = \frac{f_i^p(s_j^i)}{f_i^n(s_j^i)}, \tag{12}$$

where $f_i^p(s)$ is the conditional probability density function (PDF) of the expression value of each semantic attribute under positive category, and $f_i^n(s)$ is the conditional PDF under negative category. The ratio $\lambda_i(s_j^i)$ is a probabilistic indicator that tells us which category is more likely based on the expression value s_j^i of the i-th member attribute.

We combine the evidence from all the member attributes to infer the aggregated value of $g_j^k = \frac{1}{|g_j^k|} \sum_{i=1}^{n} \lambda_i(s_j^i)$. The proposed approach is similar to the method of computing the relative support for the two different categories based on a naive Bayes model.

In order to compute the likelihood ratio value $\lambda_i(s_j^i)$ (see Fig. 3), we need to estimate the PDF $f_i^c(s)$ for each category. We assume that the expression level of semantic attributes under category c follows the Gaussian distribution with the mean and the standard deviation, μ_i^c and σ_i^c, respectively. These parameters are estimated based on all retrieval images that correspond to the category c. The estimated PDFs can then be used for computing the likelihood ratio. In general, we often do not have enough training image set for a reliable estimation of the PDFs $f_i^p(s)$ and $f_i^n(s)$. This may make the computation of the likelihood

ratios sensitive to small changes in the scene images. To alleviate this problem, we normalize the $\lambda_i(s_j^i)$ as follows.

$$\hat{\lambda_i}(s_j^i) = \frac{\lambda_i(s_j^i) - \mu(\lambda_i)}{\sigma(\lambda_i)}, \tag{13}$$

where $\mu(\lambda_i)$ and $\sigma(\lambda_i)$ are the mean and standard deviation of $\lambda_i(s_j^i)$ across all images, respectively.

5 Experimental Result

5.1 Dataset

To evaluate the performance of the proposed method, three popular datasets are tested, *i.e.*, Scene-15 [14], Sun-15 [28], and UIUC-sports dataset [18].

Scene-15 dataset [14] consists of 15 different scene categories. Each category consists of 200 to 400 grayscale images. The images are collected from the Google image search, the COREL collection, and personal photographs.

Sun-15 dataset consists of 15 different scene categories as the same with the Scene-15 dataset. We newly created this dataset from the SUN-397 dataset [28]. The SUN-397 dataset originally contains 397 scene categories. However, in this paper, we obtained the same categories only with the Scene-15 dataset from the original dataset.

UIUC-sports dataset [18] consists of 8 sports event categories. Each sports event category is organized as follows: rowing, badminton, polo, bocce, snowboarding, croquet, sailing, and rock climbing. Images are divided into easy and medium grade according to the human subject judgement.

For a fair comparison, we follow the original experimental setup applied in [10,14]. In the experiment, 100 images per category are randomly sampled as training images and remaining images are used as test images. In case of UIUC-sports dataset, 70 images per category are randomly sampled as training images, and the remaining images are used as test images. One-versus-all strategy is used because the scene classification is a multi-class problem and the evaluated performance is reported as the average classification rate on the all categories.

5.2 Measuring Expressions of Semantic Attributes

In order to obtain semantic attributes from scene images, we employ two approaches. First approach is the semantic attribute (SA) proposed in [25]. For this, four different types of local image features are used in this experiments as in [25]: SIFT [20], LAB [17], Canny edge [4], and Texton filterbanks [16].

We measure expression values of 67 semantic attributes based on the SA as in [25] from each scene image : local scene attributes (*e.g.* building, street, tree), shape attributes (*e.g.* box, circle, cone), materials (*e.g.* plastic, wood, stone), and

objects (*e.g.* car, chair, bicycle). We learn a set of independent attribute classifiers using support vector machine (SVM) with Bhattacharyya kernel following the procedure in [25]. In order to measure the expression of each semantic attributes, non-negative SVM scores, which are obtained from results of a sigmoid function of the original SVM decision values, are used. The measured expression of each semantic attribute has a value between 0 and 1.

The other approach is the object bank (OB) proposed in [19]. The OB is obtained from an object filter response. The object responses are obtained by running a bunch of object filters across an image at various locations and scales by using the sliding window approach. Each filter is an object detector trained from images with similar view point. The models based on deformable part are applied for the object detector where six parts are used [11,19].

We measure expression values of 177 semantic attributes (objects) based on the OB as in [19] from each scene image. These 177 semantic attributes are determined from the popular image dataset such as ESP [1], LabelME [23], ImageNET [7], and Flickr! web site. After ranking the objects according to their frequencies in each of these datasets, the intersection set of the most frequent 1000 objects is obtained as the 177 semantic attributes. To train each of the 177 semantic attribute detectors, 100-200 images and their object bounding box information from the ImageNet dataset are used. In the training, a generalization of SVM called latent variable SVM (LSVM) is used for the semantic attribute detector [11]. In order to measure the expression of each semantic attribute, we follow the same procedure as in the SA case.

5.3 Discriminative Model Based Scene Classification

For scene classification, we use a SVM classifier with exponential chi-square kernel which is well-known to be suitable for the histogram-based image classification. The exponential chi-square kernel can be obtained as follows.

$$k_{chi-square}(x, y) = exp\left(-\frac{\gamma}{2}\sum_{i=1}^{n}\frac{(x_i - y_i)^2}{x_i + y_i}\right), \tag{14}$$

where γ is a scaling parameter.

The SVM-based scene classifier implemented by using LIBSVM [5]. One-versus-all strategy is used for a multi-class classification and the evaluated performance is reported as the average classification rate on the all categories. The final classification performance was obtained by average result of 50 times repeated experiments.

5.4 Comparative Results

Fig. 4 shows the performance evaluation results of the proposed methods compared with the existing methods in each semantic attribute approach. (Each semantic attribute approach is referred to as SA and OB, respectively.) As shown in Fig. 4, the proposed hypergraph-based method with the likelihood-ratio-based

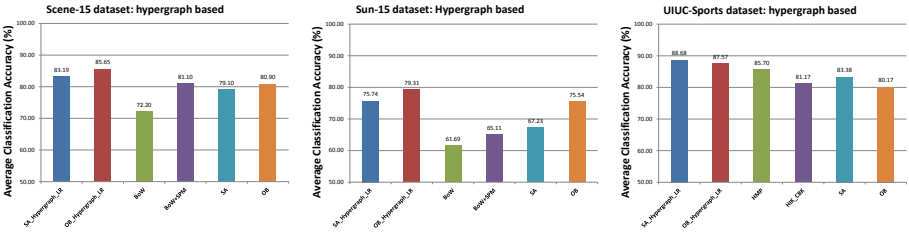

Fig. 4. Comparison of scene image classification performance with the hypergraph-based method with likelihood-ratio-based feature vector: (a) Scene-15 dataset (b) Sun-15 dataset (c) UIUC-Sports dataset

feature vector achieved outstanding performance compared with the existing methods on Scene-15, Sun-15, and UIUC-sports dataset.

In Scene-15 dataset experiment, the BoW model showed an improved performance when combined with spatial pyramid method (SPM) [14]. The methods marked as SA [25] and OB [19] are semantic-attribute-based scene classification approach in which each attribute considered individually. In the experiment, these methods showed better classification performance than the existing BoW model. However they showed lower classification result than the BoW+SPM model. On the other hand, the proposed method showed an improved result more than 4.5% compared to the results of the SA and OB considering each semantic attribute individually, even though it was not combined with the SPM unlike those in the original experiments [25,19].

In Sun-15 dataset experiment, the proposed method with the SA showed an improved result more than 10.6% compared to the result of the BoW+SPM. Especially, the proposed method with the OB showed a significantly improved result more than 14.2% compared to the result of the BoW+SPM. We can see that the method using only the Object Bank method also showed greatly improved result than existing methods. This result means that it is very important what semantic attributes are used for scene classification and how the semantic attributes are obtained from scene images.

In UIUC-sports dataset experiment, the proposed method showed an improved result more than 2.98% and 7.51% compared to the results of the HMP [2] and HIK-CBK [27], respectively. Also, in comparison with the result considering each semantic attribute individually, the proposed method showed a better result. Interestingly, unlike the previous experiments, we can see that the results using the SA-based feature vector showed better performance compared to the OB-based feature vector. We can analyze the results based on the discriminative power evaluation of feature vector. As shown in Fig. 5 (c) and Fig. 5 (f), the discriminative power of the generated feature vector based on the SA is more powerful than the generated feature vector based on the OB. More detail, the average discriminative power of top 10 features based on the SA (17.21) is bigger than the average discriminative power of top 10 features based on the OB

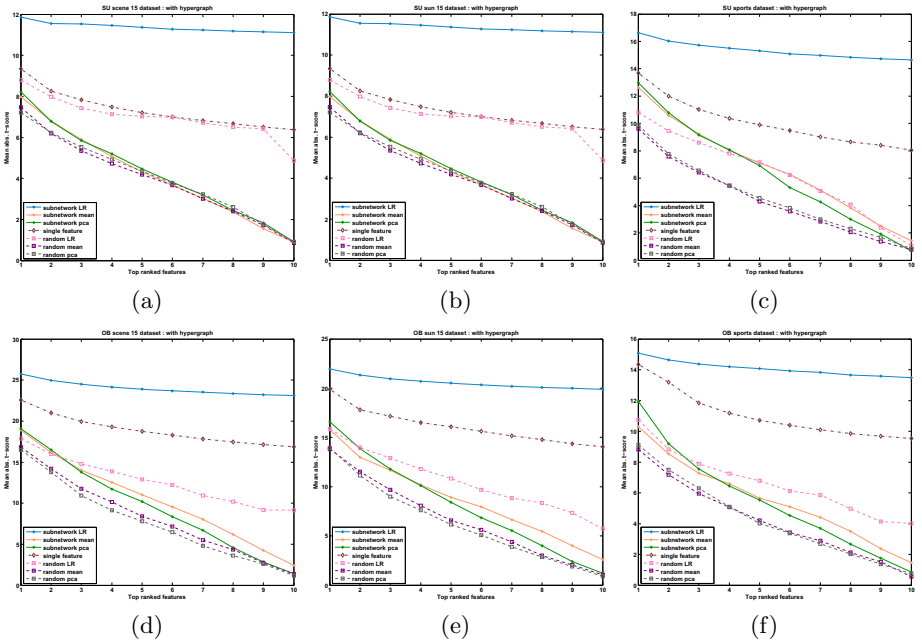

Fig. 5. Discriminative power comparison of the generated feature vectors from the SA- and OB-based semantic attributes subnetworks via the hypergraph-based method with likelihood ratio: (a) Scene-15 dataset (SA) (b) Sun-15 dataset (SA) (c) UIUC-Sports dataset (SA) (d) Scene-15 dataset (OB) (e) Sun-15 dataset (OB) (f) UIUC-Sports dataset (OB)

(15.01). Therefore, we can infer that the discriminative power of the feature vector may affect the classification performance.

Interestingly, in all experiments, we can see that the existing methods combined with the SPM were improved. We can analyze that it is a result of utilizing spatial context information. Therefore, we can expect to be able to achieve more improved performance when the subnetwork employed our method is combined with the SPM.

Previously, we analyzed that the reason why the proposed method can obtain competitive classification performance is that the discriminative power of the created feature vector by our method is strong. In order to verify this, we measured the discriminative power of feature vectors created by each method. In Fig. 5, the x-axis means rank of feature, y-axis means average absolute t-score of each feature. The discriminative power was measured using a mean absolute t-score of the top 10 features.

For a comparison, we also demonstrated a discriminative power of single semantic attribute and a discriminative power of feature vector generated from randomly selected subsets. (They are marked with single and random, respectively.) Furthermore, we compared experimental results for analyzing any difference due to the aggregation method of expression values of member attributes

constituting the subnetwork. These are marked with 'mean' and 'pca' respectively. And the proposed method is marked with 'LR'. The 'mean' method is simply averaging the expression values while the 'pca' method uses the 1st principal component of the expression values.

As shown in Fig. 5, our method not only showed strong discriminative power compared to single feature and other aggregation methods on all datasets, but also showed a tendency to maintain the strong discriminative power even in low-rank. (These results were obtained by averaging the measured results of the discriminative power of each top ranked feature for the entire category.)

This enhancement of discriminative power is related to improvement of classification performance. In addition, another advantage of our method is that it gives significantly improved performance despite the dimension of feature vector is decreased by aggregating process.

6 Conclusion

In this paper, we proposed a method of the hypergraph-based modeling, which considered the higher-order interactions of semantic attributes of a scene and applied it to a scene classification. In order to generate the hypergraph optimized for specific scene category, we proposed a novel learning method based on a probabilistic subnetworks searching and also proposed a method to generate a aggregated feature vector from the expression values of the member semantic attributes that belongs to the searched subnetworks via likelihood-based estimation.

To verify the competitiveness of the proposed method, we showed that the discrimination power of the feature vector generated by the proposed method was better than existing methods through experiments. Also, in scene classification experime9nt, the proposed method showed an outstanding classification performance compared with the conventional methods. Thus, we could regard that the consider of the higher-order interaction of the semantic attributes may have an affect on the improvement of the scene classification performance.

Acknowledgement. This research was supported by NAVER Labs and Inha University Research Grant.

References

1. von Ahn, L.: Games with a purpose. Computer 39(6), 92–94 (2006)
2. Bo, L., Ren, X., Fox, D.: Hierarchical Matching Pursuit for Image Classification: Architecture and Fast Algorithms. MIT Press (2011)
3. Bosch, A., Zisserman, A., Muñoz, X.: Scene classification using a hybrid Generative/Discriminative approach. IEEE Trans. on Pattern Analysis and Machine Intelligence 30(4), 712–727 (2008)
4. Canny, J.: A computational approach to edge detection. IEEE Trans. on Pattern Analysis and Machine Intelligence 8(6), 679–698 (1986)

5. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. on Intelligent Systems and Technology 2(3), 27:1–27:27 (2011)
6. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Proc. of ECCV Workshop on Statistical Learning in Computer Vision, pp. 1–22 (May 2004)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (June 2009)
8. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology 3(2), 185–205 (2005)
9. Everingham, M., Van Gool, L., Williams, C., Winn, J., Zisserman, A.: The PASCAL visual object classes challenge 2007 (VOC 2007) results (2007), http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/
10. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proc. of IEEE International Conference on Computer Vision, vol. 2, pp. 524–531 (October 2005)
11. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (June 2008)
12. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Tech. Rep. 7694 (March 2007)
13. Kleinberg, E.M.: Stochastic discrimination. Annals of Mathematics and Artificial Intelligence 1, 207–239 (1990)
14. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2169–2178 (June 2006)
15. Leung, T., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons. International Journal of Computer Vision 43(1), 29–44 (2001)
16. Leung, T., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons. International Journal of Computer Vision 43(1), 29–44 (2001)
17. Lew, M.S.: Principles of visual information retrieval. Springer, London (2001)
18. Li, L.J., Fei-Fei, L.: What, where and who? classifying event by scene and object recognition. In: Proc. of IEEE International Conference on Computer Vision, pp. 1–8 (October 2007)
19. Li, L.J., Su, H., Xing, E., Fei-Fei, L.: Object bank: A high-level image representation for scene classification and semantic feature sparsification. In: Advances in Neural Information Processing Systems, pp. 1378–1386. MIT Press (2010)
20. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
21. Lu, Y., Liu, P.Y., Xiao, P., Deng, H.W.: Hotelling's t2 multivariate profiling for detecting differential expression in microarrays. Bioinformatics 21(14), 3105–3113 (2005)
22. Rasiwasia, N., Vasconcelos, N.: Holistic context models for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(5), 902–917 (2012)
23. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: LabelMe: a database and web-based tool for image annotation. International Journal of Computer Vision 77(1-3), 157–173 (2008)

24. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proc. of IEEE International Conference on Computer Vision, vol. 2, pp. 1470–1477 (October 2003)
25. Su, Y., Jurie, F.: Improving image classification using semantic attributes. International Journal of Computer Vision 100(1), 59–77 (2012)
26. Voloshin, V.I.: Introduction to graph and hypergraph theory. Nova Science Publishers, Hauppauge (2009)
27. Wu, J., Rehg, J.: Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In: Proc. of IEEE International Conference on Computer Vision, pp. 630–637 (September 2009)
28. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: SUN database: Large-scale scene recognition from abbey to zoo. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3485–3492 (2010)
29. Zhang, B.T.: Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory. IEEE Computational Intelligence Magazine 3(3), 49–63 (2008)