Two-layered Photo Classification Based on Semantic and Syntactic Features

Seungji Yang, Yong Man Ro*

Image and Video Systems Lab., Information and Communications Univ.,
Munji 103-6, Yuseong, Daejeon, 305-714, South Korea
* Correspondence: yro@icu.ac.kr

Abstract. A novel approach to semantic classification for generic home photos is proposed. The proposed method consists of two-layered SVM classifiers. The first layer aims to predict the likelihood of pre-defined local photo semantics based on camera metadata and regional low-level visual features. In the second layer, one or more global photo semantics are detected based on the likelihood ratio. To construct classifiers in the first layer producing a posterior probability, we use parametric model to fit the output confidence value of SVM classifiers to posterior probability. We also exploit concept merging process based on a set of semantic-confidence map in order to cope with selecting the more likelihood photo semantics on overlapping local photo regions.

Keywords: Photo album; Semantic classification; Camera metadata; SVM

1 Introduction

Recently, it is affordable to keep a complete digital record of one's whole life. One main issue is to minimize user's manual tasks in organizing and managing a large number of photo collections. Semantic classification of arbitrary image has been a challenge in recent years. The goal of semantic classification is to discover image semantics from given pre-defined semantic concepts. The need for semantic classification has been rightly raised in digital home photo area.

One state-of-art classification approach is to use support vector machine (SVM) [13]. So far, many classification methods have employed empirical risk minimization (ERM) for learning classifier. ERM only utilizes the loss function defined for classifier and is equivalent to Bayesian decision theory with a particular choice of prior. Thus, ERM approaches often lead a classifier to be over-fitted, i.e., classifier is usually too much fitted to only training data. Unlike ERM, structural risk minimization (SRM) aims to minimize generalization error. SVM is based on the idea of SRM. The generalization error is bounded by the sum of the training set error and a term depending on the VC dimension of the learning machine. By minimizing the upper bound, high generalization can be archived. The generalization error of SVM is related not to the input dimensionality of the problem, but to the margin with separating the data. This explains why SVM can have good performance even in

problems with a large number of inputs. To date, SVM has been applied successfully to a wide range of problems.

In particular, the semantic classification problem can be usually simpler and thus easier by using multi-layered approach. Multi-layered classification approach aims to solve a classical image understanding problem that requires the effective interaction of high-level image semantics and low-level image features. Many researchers have successfully employed the multi-layered approach to semantic classification. Unfortunately, naïve SVM is inappropriate for multi-layered classifier because the output of the SVM should be a calibrated posterior probability to enable post-processing. Basically, SVM is a discriminative classifier, not based on any generative model. So, the output confidence of any classifier in a certain layer should be probabilistically modeled before being used as the probabilistic input of any classifier in the next layer. A few studies have been pressed to solve this problem [1], [2]. Platt proposed a good parametric model to fit the SVM output to the posterior probability, instead of estimating the class-conditional density. The parameters of the model are adapted to give the best probability output [1]. Lin et al. improved implementation of Platt's model [2]. They solved the problem that Platt's implementation may not converge to the minimum solution. Although Lin's method increases complexity, it gives better convergence properties.

Nevertheless, capturing high-level image semantics with low-level features remain a challenge to real application due to low performance. Unlike image, photo usually includes its camera metadata as well as pixel data itself. The metadata is obtained from Exif header from photo file [3]. Camera metadata is of great benefit to semantic photo classification in that it provides several useful cues. In particular, taken date/time stamp has been successfully employed to cluster a sequence of unlabeled photos by meaningful event or situation groups [4], [5], [6]. Especially in [4] and [5], taken date/time stamp and color features have been combined together to cluster photos by events in an automatic manner. In general, user demand for event clustering tends to exhibit little coherence in terms of low-level features, though syntactic information, such as camera metadata, could help to organize event clusters in more semantically meaningful groups. In our previous studies [7], [8], we also developed an unsupervised photo clustering scheme based on situation – that presents similar background scenery taken in a close proximity of time – as associating camera metadata and low-level features.

Especially for semantic photo classification, Boutell et al. proposed a probabilistic approach to incorporate camera metadata with content-based visual features in scene classification [9]. They exploited a useful set of camera metadata, which is related to scene brightness, flash, subject distance and focal length and verified it in some global visual semantics such as indoor/outdoor, sunset, and man-made/natural scenes. However, Boutell's method has one major disadvantage on the applications to generic scene classification. One is that, as assumed in his study, Boutell's method has limited application to a few global scenes since it used only global features, such as camera metadata and global visual features. A photo usually contains many local semantics. So, to extend the use of camera metadata to the classification of many other local and global visual semantics, the camera metadata probably need to be incorporated with visual features of local photo region. For example, let see a photo that contains human face in foreground behind background scenery. If its camera focus is on the person,

subject distance and focal length will be short. Given this knowledge, Boutell's classifier may have a difficulty of detecting background scenery in spite of using low-level visual features.

In this paper, a semantic classification scheme for generic home photos is proposed. The proposed method consists of two-layered SVM classifiers. The first layer aims to predict the likelihood of pre-defined local photo semantics based on camera metadata and regional low-level visual features. In the second layer, we determine one or more global photo semantics based on the likelihood ratio. To construct classifiers in the first layer producing a posterior probability, we use parametric model to fit the output confidence value of SVM classifiers to posterior probability. Local photo semantics provide an intermediate level of photo semantics by bridging the semantic gap of low-level features and high-level photo semantics. We also exploit concept merging based on a set of semantic-confidence map so as to cope with selecting the more likelihood photo semantics on overlapping local photo regions. For multi-class determination in global photo semantics, we propose to use three different criterions.

2 Method

2.1 Local Semantic Classification

2.1.1 Regional Division for Local Semantics

Most of the current digital cameras support auto-focusing (AF) system that works as moving the camera lens in and out until the sharpest possible image of the subjects is projected onto the image receptor such as CCD and CMOS. All AF systems provide a certain number of censoring regions. A censoring region usually forms rectangle. This means that photographer's intension can be found in the rectangle censoring regions.

Indeed, the best representation of local visual semantics in photo is given by object segmentation, which could produce elaborate object contours. So far, however, there seems no almighty method for object segmentation. Rather, the object segmentation is usually expensive in computation and even sometimes produces incomplete results in complex natural images. So, instead, we approach a simple block segmentation to capture visual semantics that appear on local photo regions. The block segmentation is relatively inexpensive. But, to boost its low segmentation performance, we employ a set of region template, denoted as photographic region template (PRT), whose idea originates from the rectangle censoring system of digital camera. Thus, although PRT is used in a block tessellation with a fixed number of blocks, it could be fast and good enough to detect what the photographer intended to capture when taking the picture. The basic observation behind the PRT is that mainly-concerned subjects would be usually focused, taking larger portion and being sharpener than other un-concerned subjects. Thus, many other most likely small, blurred subjects would be often out of concern in the photo.

In order to build meaningful region templates, three conditions are considered: the region template should be large enough to detect semantics in the local photo region, simultaneously be small enough not to be time-consuming in feature extraction and similarity measure, and support spatial scalability to detect photo semantics over various scale subjects. From this observation, we propose a photographic region template as shown in Fig. 1. The region template is composed of ten local regions: one center region (R_1 in Fig. 1), four corner regions (R_2, R_3, R_4, and R_5 in Fig. 1), two horizontal regions (R_6 and R_7 in Fig. 1), two vertical regions (R_8 and R_9 in Fig. 1), and a whole photo region (R_{10} in Fig. 1). The four corner regions are parts of the vertical, horizontal, and whole regions. Note that one center and four corner regions are referred to as basis regions. The use of basis region set will be presented in local semantic classification. The center region overlaps partially with the corner, vertical and horizontal regions, and entirely with the whole photo region.

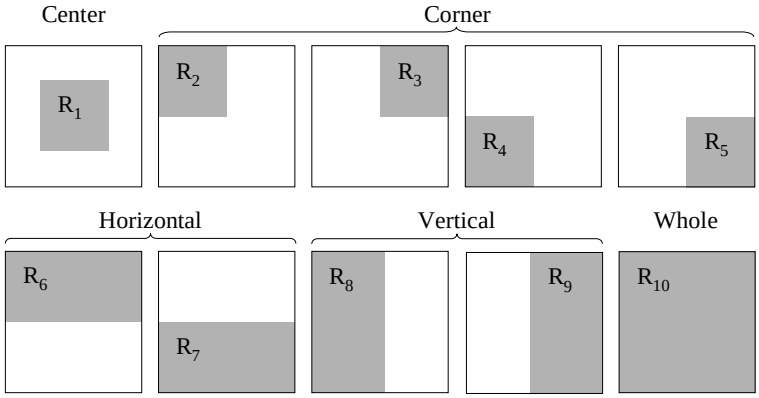

Fig. 1. Photographic region templates

2.1.3 Local Semantic Learning

SVM is employed as local semantic classifiers in the first layer. It gives a good binary classifier that is used to find the decision function of optimal linear hyper-plane given labeled training data. SVM is a constructive learning procedure rooted in statistical learning theory [13]. It is based on the principle of structural risk minimization, which aims at minimizing the bound on the generalization error rather than minimizing the mean square error over the data set. As a result, an SVM tends to perform well when applied to data outside the training set. The hyper-plane can be linearly separable in high-dimensional feature space (h). Input feature in the space ($\mathbf{F}$) is mapped onto the feature space via a nonlinear mapping ($\varphi: \mathbf{F} \to h$), allowing one to perform nonlinear analysis of the input features using a linear method. In generic SVM, a kernel is designed to map the input data space to the feature space. With the 'kernel trick' property [10], the kernel can be considered as similarity measures between two feature vectors without explicit computation of the map φ. Using kernel function, SVM classifier can be trained with features of training data. For this, an optimal hyper-plane is found to correctly classify the training data. By the optimization

theorem of SVM, the decision function (Φ_n^{local}) to predict the local concept (x_n^{local}) of unseen feature vector ($\mathbf{F}$) is formed as follows,

$$\Phi_n^{local}(\mathbf{F}) = \sum_t a_n^t z_n^t K(\mathbf{F}_n^t, \mathbf{F}) + b_n , \quad (\mathbf{1})$$

where K is a kernel function that can be a linear function, radial-basis function (RBF), polynomial function, sigmoid function, etc., and, in this paper, RBF kernel fuction that is the most popular choice of kernel types is selected. $\mathbf{F}_n^t$ is the t[th] support vector of the hyper-plane for the local concept (x_n^{local}), a_n is the vector of corresponding weighting values of the support vector, z_n is the corresponding class vector of the support vector, and b_n is the threshold optimized for the local concept (x_n^{local}).

Constructing the SVM classifier to produce a posterior probability, the output confidence value of the SVM is fitted to a parametric sigmoid model [1, 2]. The form of parametric sigmoid fitting model for the classifier of a local photo semantic x_n^{local} is as follows,

$$P_n(y = 1 \mid \Phi_n^{local}(\mathbf{F})) = \frac{1}{1 + \exp(A \cdot \Phi_n^{local}(\mathbf{F}) + B)} , \quad (\mathbf{2})$$

where A and B are parameters to determine the shape of the sigmoid model. So, the SVM output ranged from $-\infty$ to ∞ is fitted to the probabilistic output ranged from 0 to 1.

The best parameters (A, B) are estimated by solving the following regularized maximum likelihood problem with a set of labeled training example. Given a training set $(\Phi_n^{local}(\mathbf{F}_i), y_i)$, let us define a new training set $(\Phi_n^{local}(\mathbf{F}_i), y'_i)$, where the y'_i is target probability value. The new target value is used instead of (0, 1) for all of the training data in the sigmoid fit. This aims at making the new target value converge to (0, 1) when the training set size approaches infinity. The new target value y'_i is defined as follows,

$$y'_i = \begin{cases} \dfrac{N_+ + 1}{N_+ + 2}, & y_i = 1 \\ \dfrac{1}{N_- + 2}, & y_i = -1 \end{cases} , \quad (\mathbf{3})$$

where N_+ is the number of positive samples and N_- is the number of negative samples. Then, the best parameters for a local photo semantic are obtained as minimizing the following cross-entropy error function.

$$\arg\min_{(A,B)} -\sum_i \{y'_i \cdot \log p_i + (1 - y'_i)\log(1 - p_i)\}, \quad (\mathbf{4})$$

where p_i denotes $P_n\left(y_i \mid \Phi_n^{local}(\mathbf{F})\right)$. We adopt Lin's method [2] to find the optimized parameters minimizing the above error function.

2.1.3 Integration of Camera Metadata and Local Visual Features

To integrate camera metadata with low-level visual features in the proposed photo classification, we first generalize the following probabilistic combination scheme. Let $\mathbf{X}=\{x_1, x_2, \cdots, x_I\}$ be a set of I photo semantic classes that frequently appear in home photos. And, let $\mathbf{F}_{cam}=\{f_{cam}^1, f_{cam}^2, \cdots, f_{cam}^J\}$ be a useful set of J camera metadata, and $\mathbf{F}_{low}=\{f_{low}^1, f_{low}^2, \cdots, f_{low}^K\}$ be a set of K low-level visual features. Then, the likelihood of a semantic class, $x_i \in \mathbf{X}$, on the given features, $\mathbf{F}=\{\mathbf{F}_{cam}, \mathbf{F}_{low}\}$, can be represented by the joint conditional probability as follows,

$$P(x_i|\mathbf{F}) = P(x_i|\mathbf{F}_{cam}, \mathbf{F}_{low}), \quad (5)$$

By the Bayesian theorem, the joint conditional probability can be decomposed as follows,

$$P(x_i|\mathbf{F}) = P(x_i|\mathbf{F}_{cam}, \mathbf{F}_{low}) = \frac{P(x_i)P(\mathbf{F}_{cam}, \mathbf{F}_{low}|x_i)}{P(\mathbf{F}_{cam}, \mathbf{F}_{low})}, \quad (6)$$

Let us embody (1) to local semantics. For this, let $\mathbf{X}^{local}=\{x_1^{local}, x_2^{local}, \cdots, x_N^{local}\}$ be a set of N local semantics. Then, the joint conditional probability of a local semantic $x_n^{local} \in \mathbf{X}^{local}$ given an input feature set $\mathbf{F}^{local}=\{\mathbf{F}_{cam}, \mathbf{F}_{low}^{local}\}$ – where camera metadata is not local, but global – for the local photo regions can be written as follows,

$$P(x_n^{local}|\mathbf{F}^{local}) = P(x_n^{local}|\mathbf{F}_{cam}, \mathbf{F}_{low}^{local}) = \frac{P(x_n^{local})P(\mathbf{F}_{cam}, \mathbf{F}_{low}^{local}|x_n^{local})}{P(\mathbf{F}_{cam}, \mathbf{F}_{low}^{local})}, \quad (7)$$

The camera metadata ($\mathbf{F}_{cam}$) is independent of the low-level features ($\mathbf{F}_{low}^{local}$), so that (3) can be written again as follows,

$$\frac{P(x_n^{local})P(\mathbf{F}_{cam}, \mathbf{F}_{low}^{local}|x_n^{local})}{P(\mathbf{F}_{cam}, \mathbf{F}_{low}^{local})} = \frac{P(x_n^{local})P(\mathbf{F}_{cam}|x_{low}^{local})P(\mathbf{F}_{cam}|x_n^{local})}{P(\mathbf{F}_{cam})P(\mathbf{F}_{low}^{local})}, \quad (8)$$

2.1.4. Local Semantic Classification

As mentioned above, the input photo to be classified is divided into ten local regions by the photographic region template. Multiple low-level visual features are extracted from each local region and fed into the local concept detectors. For the local photo semantic classification, let $\mathbf{R}=\{R_1, R_2, \cdots, R_{10}\}$ be a set of the local regions. Then, the feature vector of a local region ($R \in \mathbf{R}$) is denoted as $\mathbf{F}^R=\{\mathbf{F}_{cam}, \mathbf{F}_{low}^R\}$. Equations (7) and (8) can be specified for the local region as follows,

$$P\left(x_n^{local}\middle|\mathbf{F}^R\right)=P\left(x_n^{local}\middle|\mathbf{F}_{cam},\mathbf{F}_{low}^R\right)=\frac{P\left(x_n^{local}\right)P\left(\mathbf{F}_{cam}\middle|x_n^{local}\right)P\left(\mathbf{F}_{low}^R\middle|x_n^{local}\right)}{P\left(\mathbf{F}_{cam}\right)P\left(\mathbf{F}_{low}^R\right)}, \quad (\mathbf{9})$$

where the camera metadata ($\mathbf{F}_{cam}$) and corresponding probability $P\left(\mathbf{F}_{cam}\middle|x_n^{local}\right)$ is the same over all local regions given an input photo. The $P\left(\mathbf{F}_{low}^R\middle|x_n^{local}\right)$ is regarded as the probability of the local region feature ($\mathbf{F}_{low}^R$) about the SVM model of the local concept (x_n^{local}). So, it is estimated by the sigmoid model as follows,

$$P\left(\mathbf{F}_{low}^R\middle|x_n^{local}\right)\approx\frac{1}{1+\exp\left(A\cdot\Phi_n^{local}\left(\mathbf{F}_{low}^R\right)+B\right)}. \quad (\mathbf{10})$$

Similarly, the $P\left(\mathbf{F}_{cam}\middle|x_n^{local}\right)$ is regarded as the probability of the camera metadata feature ($\mathbf{F}_{cam}$) about the SVM model of the local concept (x_n^{local}).So, it is also estimated by a sigmoid function as follows,

$$P\left(\mathbf{F}_{cam}\middle|x_n^{local}\right)\approx\frac{1}{1+\exp\left(A\cdot\Phi_n\left(\mathbf{F}_{cam}\right)+B\right)}. \quad (\mathbf{11})$$

Over all local regions ($\mathbf{R}$), the probability set of the local concept (x_n^{local}) can be written as follows,

$$P\left(x_n^{local}\middle|\mathbf{F}_{low}^{\mathbf{R}}\right)=\left\{P\left(x_n^{local}\middle|\mathbf{F}_{low}^{R_1}\right),P\left(x_n^{local}\middle|\mathbf{F}_{low}^{R_2}\right),\cdots,P\left(x_n^{local}\middle|\mathbf{F}_{low}^{R_{10}}\right)\right\}. \quad (\mathbf{12})$$

Given $\mathbf{X}^{local}=\left\{x_1^{local},x_2^{local},\cdots,x_N^{local}\right\}$, the probability set of the local concept set ($\mathbf{X}^{local}$) can be written as follows,

$$\begin{aligned}P\left(\mathbf{X}^{local}\middle|\mathbf{F}_{low}^{\mathbf{R}}\right)&=\left\{P\left(x_1^{local}\middle|\mathbf{F}_{low}^{\mathbf{R}}\right),P\left(x_2^{local}\middle|\mathbf{F}_{low}^{\mathbf{R}}\right),\cdots,P\left(x_N^{local}\middle|\mathbf{F}_{low}^{\mathbf{R}}\right)\right\}\\&=\begin{Bmatrix}P\left(x_1^{local}\middle|\mathbf{F}_{low}^{R_1}\right),P\left(x_1^{local}\middle|\mathbf{F}_{low}^{R_2}\right),\cdots,P\left(x_1^{local}\middle|\mathbf{F}_{low}^{R_{10}}\right),\cdots\\,P\left(x_N^{local}\middle|\mathbf{F}_{low}^{R_1}\right),P\left(x_N^{local}\middle|\mathbf{F}_{low}^{R_2}\right),\cdots,P\left(x_N^{local}\middle|\mathbf{F}_{low}^{R_{10}}\right)\end{Bmatrix},\end{aligned} \quad (\mathbf{13})$$

If $v_{n,R}^{local}=P\left(x_n^{local}\middle|\mathbf{F}_{low}^R\right)$, (12) can be written again as follows,

$$\mathbf{V}^{local}=\left\{v_{1,1}^{local},v_{2,1}^{local},\cdots,v_{n,1}^{local},v_{1,2}^{local},v_{2,2}^{local},\cdots,v_{n,2}^{local},\cdots,v_{1,10}^{local},v_{2,10}^{local},\cdots,v_{n,10}^{local}\right\} \quad (\mathbf{14})$$

where $v_{n,R}^{local}$ stands for the degree of likelihood of the n local concept set about the R local regions feature. Table 1 shows the probability of the local concept for each local region.

2.2 Global Semantic Classification

2.2.1 Association of Local Semantics with Global Semantics

We express the degree of strength of the semantic link between local semantics and global semantics. The higher value stands for a stronger connection between concepts. This approach could bridge the semantic gap between low-level features and high-level concepts. Thus, the global concepts are trained based on the confidence vectors of the local SVM models. Similar to the local concepts, the decision function (Φ_m^{global}) to predict the local concept (x_m^{global}) of unseen confidence feature vector ($\mathbf{V}_R^{local}$) given local regions ($\mathbf{R}$) is formed as follows,

$$\Phi_m^{global}\left(\mathbf{V}^{local}\right) = \sum_t a_m^t z_m^t K\left(\mathbf{V}_m^t, \mathbf{V}^{local}\right) + b_m \ , \tag{15}$$

where $\mathbf{V}_m^t$ is support vector of the hyper-plane for the global concept (x_m^{global}).

To find more likelihood semantics on the overlapping local regions, a concept merging is performed by keeping the most confident concepts for the five basis local regions ($\mathbf{R}^{basis}$) that consists of one center and four corner regions, that is, the region set can be defined as $\mathbf{R}^{basis} = \{R_1, R_2, R_3, R_4, R_5\}$, where rightly $\mathbf{R}^{basis} \subset \mathbf{R}$. The concept merging is performed with semantic confidence map used to keep the most confident concept for the basis local regions set.

The semantic confidence map gives five different combinations of overlapping local regions as shown in Fig. 3. Then, the confidence value of a local concept (x_n^{local}) of the a basis region ($R_b \in \mathbf{R}^{basis}$) is calculated as follows,

$$v_{n,b}^{local} = \max\left(v_{n,t}^{local} \middle| t \in \mathbf{R}_b^{map}\right), \tag{16}$$

where, for example, if the basis region is R_2, $v_{n,2}^{local} = \max\left(v_{n,t}^{local} \middle| t \in \{2,6,8,10\}\right)$.

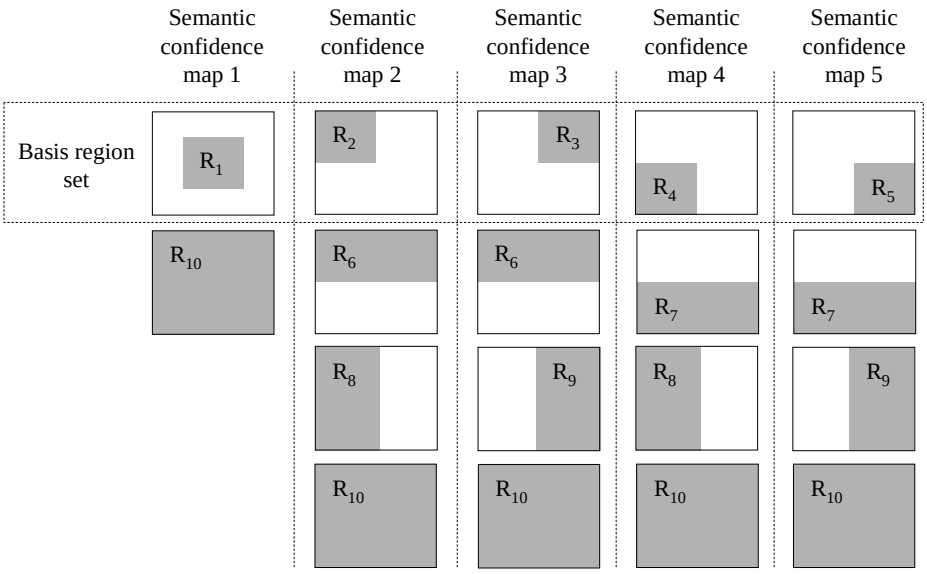

Fig. 2. Semantic confidence map

2.2.2 Global Semantic Classification

Given a basis local region, the merged confidence values for all local concepts are used to classify the local regions into the target classes. In this paper, one of the main targets is to detect multi-classes, meaning that an input photo can be labeled by one or more classes. For this, we propose three criterions for multi-class categorization. Given the probability values for the five basis local regions of an input photo, the three categorization criterions are as follows:

1) α criterion: In this case, every basis local regions can have only one class whose probability value is the top-most over all global concept classes given each basis local region.

2) β criterion: In this case, every basis local regions can have only one or no class. That is, a basis local region can have a single class whose probability value is close enough, i.e., higher than a threshold.

3) γ criterion: In this case, first of all, the probability values for all basis local region are aligned in ascending order. Then, the top-N classes with respect to the probability value are assigned to classes of the input photo, whose probability values should be close enough, i.e., higher than a threshold.

In the case of α criterion, the classifier assigns the class of a basis local regions (R_b) to a concept satisfying the following MAP condition, given by,

$$c_\alpha = \arg\max_{c=1,2,\cdots,M}\left\{\frac{P\left(x_c^{global}\right)\prod_{t=1}^{N}P\left(v_{t,b}^{local}\left|x_c^{global}\right.\right)}{\prod_{t=1}^{N}P\left(v_{t,b}^{local}\right)}\right\} = \arg\max_{c=1,2,\cdots,M}\left\{P\left(x_c^{global}\right)\prod_{t=1}^{N}P\left(v_{t,b}^{local}\left|x_c^{global}\right.\right)\right\}, \quad (17)$$

where c_α is one predicted class of the basis local regions. Accordingly, the classifier by α criterion generates five predicted classes for an input photo.

In the case of β criterion, the classifier assigns the class of a basis local regions (R_b) to one concept or none satisfying the following condition, given by,

$$c_\beta = \begin{cases} c_\alpha, & \text{if } P\left(x_{c_\alpha}^{global}\right)\prod_{t=1}^{N}P\left(v_{t,b}^{local}\left|x_{c_\alpha}^{global}\right.\right) \geq P_{th} \\ none, & otherwise \end{cases}, \quad (18)$$

where c_β is the predicted class of the basis local regions, and P_{th} is the threshold value for categorization criterion. Accordingly, the classifier by β criterion generates five or less than five predicted classes for input photos.

In the case of γ criterion, the classifier assigns the class of an input photo to multiple concepts satisfying the following condition, given by,

$$c_\gamma = c, \text{if } P\left(x_c^{global}\right)\prod_{t=1}^{N}P\left(v_{t,b}^{local}\left|x_c^{global}\right.\right) \geq P_{th} \text{ for any class and any basis region}, \quad (19)$$

where c_γ is the predicted class of the input photo.

3 Experiments

To demonstrate the proposed photo classification, experiments were performed with the official database of the MPEG-7 visual core experiment 2 (VCE-2) test data set that comprises 3086 real home photos. The goal of the MPEG-7 VCE-2 is to verify the usefulness of the MPEG-7 visual descriptors for photo classification. All of the photos in the database were contributed by several participants in the MPEG-7 VCE-2. The MPEG-7 VCE-2 also provides corresponding ground truth (GT) set for the databases.

The official GT set is given by seven semantic classes that would popularly appear in home photos. It was cross-verified by several participants in the MPEG-7 VCE-2 who are experts in content-based image analysis. The seven semantic classes includes ‘architecture’, ‘indoor’, terrain’, ‘night’, ‘snowscape’, ‘waterside’, and ‘sunset’. Note that the GT set was strictly made to avoid missing any human visual preference in browsing photos. That is, an important rule in the GT decision was that a photo could be labeled with one or more semantic classes of which a scene could be detectable by the human eye. Therefore, many of the photos were labeled by multiple classes.

As totally independent of the test data set, 1597 photos were used for training data. They were also from the MPEG-7 VCE-2 official training data set. Of the training set,

800 were from general home photos, and 797 were from the Corel photo collection. For training local semantic classifier, we patched the training photos to local regions and then manually selected positive and negative samples for each class from the sub-photo collection by human visual perception. The negative samples for each concept were randomly selected from the positive samples of other all concepts.

For learning local semantics, multiple low-level visual features are extracted from the patched photo database. For this, five MPEG-7 descriptors are employed for color and texture features [11], [12]: color structure (CS), color layout (CL), and scalable color (SC) descriptors are used for color features; and homogeneous texture (HT) and edge histogram (EH) descriptors are used for texture features.

In this paper, we build nine important families of concepts that would frequently appear in local regions of general home photos. The families of the local concepts consists of 'ground', 'human', 'indoor', 'mountain', 'night', 'plant', 'sky', 'structure', and 'water'. The concept families are sub-divided to the 34 local concepts as follows:

- Seven 'ground' concepts: 'gravel', 'park', 'pavement', 'road', 'rock', 'sand', and 'sidewalk';
- Two 'human' concepts: 'face' and 'people';
- Two 'indoor' concepts: 'indoor' and 'indoor-light';
- Three 'mountain' concepts: 'field', 'peak', and 'wood';
- Two 'night' concepts: 'night' and 'street-light';
- Three 'plant' concepts: 'flowers', 'leaves', and 'trees';
- Four 'sky' concepts: 'cloudy', 'sunny', 'sunset', and 'sunset-on-mountain';
- Five 'structure' concepts: 'brick', 'arch', 'buildings', 'wall', and 'windows';
- Six 'water' concepts: 'beach', 'high-wave', 'low-wave', 'still water', 'mirrored water', and 'ice (snow)'

Accuracy, recall, and precision are well-known measures to evaluate classification performance. As in general definition, accuracy = (TP + TN) / (total number of samples), recall = TP / (TP + FN), and precision = TP / (TP + FP), where TP, TN, FP, and FN stand for 'true positive' when the case is positive and predicted positive, 'true negative' when the case is negative and predicted negative, 'false positive' when the case is negative but predicted positive and 'false negative' when the case is positive but predicted negative, respectively.

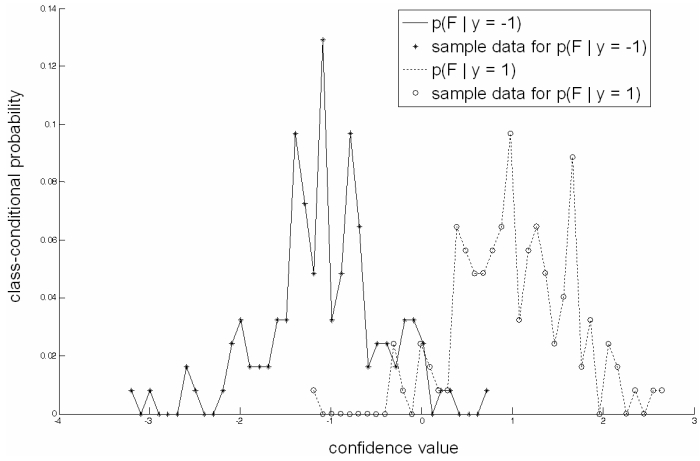

Fig. 3. The histogram of positive and negative samples for indoor classifier

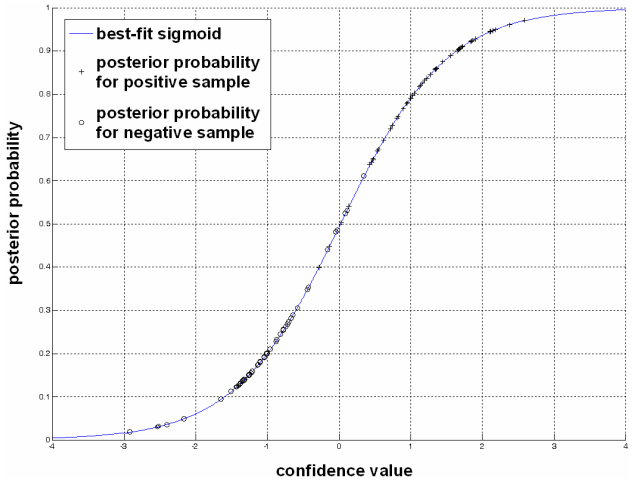

Fig. 4. The best fit sigmoid for indoor classifier

The sigmoid parameters were calculated for each local semantic classifier. Fig. 3 shows the histogram of positive and negative samples for indoor classifier. The solid line is the class-conditional probability of negative samples, while the dashed line is that of positive samples. As shown in Fig. 3, the histogram is not Gaussian, probably due to the small number of training data. Fig. 4 is derived by using Bayes' rule on the histogram estimates of the class-conditional densities. The sigmoid fit works well, as can be seen in Fig. 4.

First, we measured classification performance without local semantic features, i.e., with only global low-level features. In Table 1, (a) column shows its average performance for each global concept. The average performance was measured with a threshold showing minimum difference between recall and accuracy. The results show that night class has the best performance at about 90% and architecture class the

worst at about 61%. To verify the usefulness of the two-layered classification scheme, we also measured classification performance with local low-level features and local semantic features. In Table 1, (b) column shows its average result for each global concept. Adding local semantic features made global semantic classification perform much better in indoor class, as compared with the case of using only global low-level features. Thus, local semantic features would be useful to catch local indoor semantics. In other classes, recall and accuracy slightly increased.

Table 1. Comparison of average performances of classification methods with (a) global low-level feature, and (b) with local low-level features and local semantic features

Class	(a) Average performance with global low-level features (recall/accuracy)	(b) Average performance with local low-level features and local semantic features (recall/accuracy)
Architecture	61.05 (61.01/61.08)	62.37 (61.91/62.82)
Indoor	64.10 (63.98/64.21)	74.14 (74.35/73.93)
Terrain	67.56 (67.52/67.60)	71.36 (69.98/72.73)
Night	89.62 (89.53/89.71)	90.70 (91.55/89.84)
Snowscape	75.30 (75.76/74.84)	81.39 (81.21/81.56)
Sunset	76.94 (76.47/77.41)	79.07 (77.94/80.19)
Waterside	67.83 (67.64/68.02)	70.88 (72.12/69.94)

The camera metadata includes exposure-time (refer to ET), aperture number (refer to AN), focal length (refer to FL), and flash-fired or not (refer to FF). It is denoted that the camera metadata would be considered for only indoor/outdoor and night/daytime classes since it would not be useful for other semantic classes.

So, given this constraint, in order to employ the camera metadata in local semantic classification, we first constructed two local semantic classifiers: indoor/outdoor and night/daytime classifiers. Fig. 5 shows the two local semantic classifiers with camera metadata. Fig. 5-(a) shows the indoor/outdoor classifier that outputs probability values for indoor and outdoor classes by using several useful camera metadata as syntactic features. Similarly, Fig. 5-(b) shows the night/daytime classifier that outputs probability values for night and daytime classes by using several useful camera metadata for syntactic features. In order to associate the two classifiers with the 34 local concepts, we make a classification scheme as seen in Fig. 5-(c). As such, the first step is to classify the input camera metadata into indoor or outdoor classes. The indoor probability is assigned to indoor classes, and the outdoor probability is assigned to outdoor classes. The second step is to classify the input camera metadata into night and daytime classes. The night probability is assigned to night classes and the daytime probability is assigned to daytime classes that include ground, human, mountain, sky, structure, plant, and water classes.

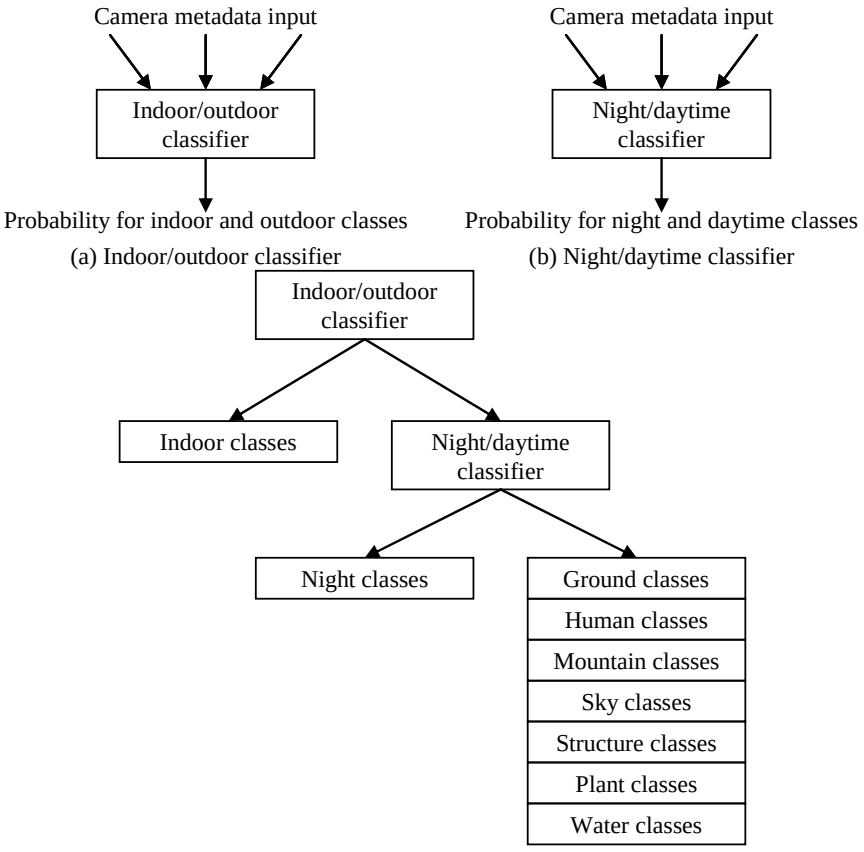

Fig. 5. Local semantic classifiers with camera metadata

Table 1 shows the performance of the proposed photo classification scheme that associates local semantic features with camera metadata. A few classes have been better classified with α or β criterions, but γ criterion has been the best over almost all classes. The proposed method, which uses local photo semantic features incorporated with both local low-level features and camera metadata, has increased the classification performance more than the method, which uses local photo semantic features incorporated with only local low-level features. There is about 5% increase in architecture class, about 10% increase in indoor class, about 2% increase in terrain class, about 1% increase in night class, about 2% increase in snowscape class, and about 1% increase in waterside class. Fig. 6 shows classification performance in the case of associating camera metadata with local low-level features and local semantic features over possible thresholds.

Table 2. Classification performances with local semantic features and camera metadata

Class	Recall (%)			Accuracy (%)		
	α	β	γ	α	β	γ
Architecture	55.63	55.63	70.13	68.74	68.74	72.23
Indoor	83.57	82.42	83.59	78.00	84.78	82.33
Terrain	31.66	31.66	80.24	62.86	62.86	77.48
Night	61.49	61.49	94.38	92.54	92.54	92.32
Snowscape	93.94	83.03	83.10	71.78	82.72	84.56
Sunset	66.18	66.18	81.67	75.39	75.39	82.54
Waterside	53.87	53.87	74.32	72.51	72.51	72.72

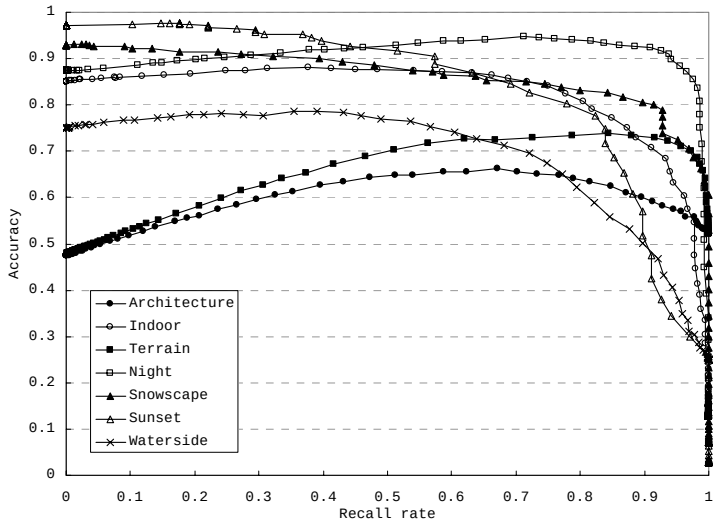

Fig. 6. Classification performances with local semantic features and camera metadata overall all thresholds: γ criterion

The proposed method was also compared with related work using Bayesian network classifier with global visual features and camera metadata [9]. The main difference of our method from the Boutell's one is that we provide a scheme to employ local semantic features especially for the two-layered SVM classifier. Our assumption is that the proposed method will outperform the conventional one in local photo semantic classification. Table 3 shows the categorization results of the two different methods. The training and testing data was the same as the above experiment. As seen in the results, almost categories except for architecture were better detected by the proposed method than by the conventional method. In indoor and terrain, both methods showed similar performance. But, the proposed method much better detected other categories such as night, snowscape, sunset and waterside.

Table 3. Classification performances with local semantic features and camera metadata

Performance / Category	Bayesian network			Proposed two-layered SVM		
	Recall	Accuracy	Average	Recall	Accuracy	Average
Architecture	87.34	70.97	79.16	70.13	72.23	71.18
Indoor	96.64	71.93	84.29	83.59	82.33	82.96
Terrain	90.28	66.32	78.30	80.24	77.48	78.86
Night	84.14	67.79	75.97	94.38	92.32	93.35
Snowscape	87.43	41.95	64.69	83.10	84.56	83.83
Sunset	84.29	58.59	71.44	81.67	82.54	82.11
Waterside	73.08	55.41	64.25	74.32	72.72	73.52

4 Conclusions

This paper exploits a scheme to employ syntactic features, such as camera metadata, for semantic classification. We select a two-layered approach to detect local and global photo semantics. The camera metadata provide useful cues independent of photo contents, facilitating the discovery of photo semantics. Our approach is characterized in the following two schemes: one is the scheme that incorporates syntactic features to low-level visual features for detecting local photo semantics; the other is the scheme that uses the local photo semantics as features for detecting global photo semantics. Concept merging is also proposed to select more likelihood semantic concepts on overlapping local regions. The efficacy of the proposed categorization method was demonstrated with 3086 MPEG-7 VCE-2 official databases. The experiment results showed that the proposed method would be useful to detect multiple semantic meaning of generic home photos. In future, we will extend the application of the proposed classification scheme to other syntactic features. In addition, we need to compare the proposed method to other similar approaches such as to Boutell's using Bayesian network.

References

1. Platt J.: Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: A.J. Smola, P.L. Bartlett, B. Sch¨olkopf, and D. Schuurmans, Advances in Large Margin Classifiers, Cambridge, MIT Press (2000)
2. Lin H.T., Lin C.J., and Weng R.C.: A note on Platt's probabilistic outputs for support vector machines. Technical report, National Taiwan University (2003)
3. Exchangeable image file format for digital still cameras, JEITA CP-3451, Japan Electronics and Information Technology Industries Association
4. Loui A.C. and Savakis A.: Automated event clustering and quality screening of consumer pictures for digital albuming. IEEE Trans. of Multimedia. 5(3) (2003) 390-402
5. Lim J.H, Tian Q., and Mulhem P.: Home photo content modeling for personalized event-based retrieval. IEEE Trans of Multimedia. 10(4) (2003) 24-37

6. Cooper M., Foote J., Girgensohn A., and Wilcox L.: Temporal event clustering for digital photo collections. Proc. of ACM Multimedia. (2003) 364-373
7. Yang S., Yoon J.H., Kang H.K., and Ro Y.M.: Category Classification using Multiple MPEG-7 Descriptors. CISST. 1 (2002) 396-401
8. Yang S., Yoon J.H., and Ro Y.M.: Automatic Image Categorization using MPEG-7 Description. Proc. of SPIE Electronic Imaging on Internet Imaging. 5018 (2003) 139-147
9. Boutell M., Luo J.: Beyond pixels: Exploiting camera metadata for photo classification. Pattern Recognition. 38 (2005) 935-946
10. Muller K.: An introduction to kernel-based learning algorithms. IEEE Trans. on Neural Networks. 12(2) (2001) 181-201
11. Ro Y.M., Kang H.K.: Hierarchical rotational invariant similarity measurement for MPEG-7 homogeneous texture descriptor. Electronics Letters. 36(15) (2000) 1268-1270
12. Yang S., Yoon J.H., and Ro Y.M.: Automatic Image Categorization using MPEG-7 Description. Proc. of SPIE Electronic Imaging on Internet Imaging. 5018 (2003) 139-147
13. Vapnik, V. N.: The Nature of Statistical Learning Theory, second ed. Springer (1999)