Hindawi
Scientific Programming
Volume 2020, Article ID 3878153, 13 pages
https://doi.org/10.1155/2020/3878153

Research Article

Multiview Active Learning for Scene Classification with High-Level Semantic-Based Hypothesis Generation

Tuozhong Yao,[1] Wenfeng Wang ,[1,2] Yuhong Gu,[3] and Qiuguo Zhu[4]

[1]*School of Electronic and Information Engineering, Ningbo University of Technology, Ningbo 315211, China*
[2]*School of Electronic and Electrical Engineering, Shanghai Institute of Technology, Shanghai 200235, China*
[3]*Shihezi Medical School, Shihezi 832000, China*
[4]*State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310007, China*

Correspondence should be addressed to Wenfeng Wang; wangwenfeng@nimte.ac.cn

Received 7 February 2020; Revised 23 April 2020; Accepted 18 August 2020; Published 1 September 2020

Academic Editor: Chenxi Huang

Multiview active learning (MVAL) is a technique which can result in a large decrease in the size of the version space than traditional active learning and has great potential applications in large-scale data analysis. This paper made research on MVAL-based scene classification for helping the computer accurately understand diverse and complex environments macroscopically, which has been widely used in many fields such as image retrieval and autonomous driving. The main contribution of this paper is that different high-level image semantics are used for replacing the traditional low-level features to generate more independent and diverse hypotheses in MVAL. First, our algorithm uses different object detectors to achieve local object responses in the scenes. Furthermore, we design a cascaded online LDA model for mining the theme semantic of an image. The experimental results demonstrate that our proposed theme modeling strategy fits the large-scale data learning, and our MVAL algorithm with both high-level semantic views can achieve significant improvement in the scene classification than traditional active learning-based algorithms.

1. Introduction

Scene classification is defined as using a computer to understand the class of an image scene. The related research studies can be roughly divided into two branches: some focus on fast holistic scene perception based on visual psychology and physiology [1, 2], while others build the statistical models through local image analysis to understand the scene, which is also the main developing tendency [3–5]. There have been many methods for image representation in the past two decades, which is a key step for scene classification. Low-level features such as color, texture, and edge have been widely used to represent the local regions of an image. Some researchers trained object detectors to achieve high-level semantics such as object's class, size, and shape for more accurate image representation [6, 7]. Prevailing statistical models are bag-of-words (BoW) and related theme statistical models. These models reduce the gap between the low-level features and high-level semantics by mining the hidden themes from local image regions such as pLSA [8] and LDA [9]. Other new scene statistical models [10–12] were proposed for more accurate object recognition in the scene. However, these mentioned models above mainly focus on the occurrence of the image semantics, and the spatial semantic correlations between different image regions are usually ignored.

For mining the spatial context information from an image, some researchers considered the information interaction between different spatial pyramid levels [13–15], and how to build reasonable attention mechanisms also can lead to significant improvement for scene classification. These methods used deep neural networks, and their large-scale network parameter estimation tasks usually lead to much higher computational complexity than nondeep learning based methods.

Active learning ranks the unlabeled samples iteratively and only selects the samples with high uncertainty or which cause great ambiguity for the classifier. In PAC learning theory, compared with traditional passive learning, it can exponentially reduce its sample complexity to $O(\log(1/\varepsilon))$ in the feature space for learning a classifier with expectation classification error ε [16–18], which has good potential of wide application in large-scale data leaning. However, most of the traditional active learning algorithms' lack of diversity of the hypotheses is generated usually by low-level image features, which affects their performances. This paper proposed a MVAL-based scene classification algorithm, which uses different high-level semantics as its views and can realize a decrease in more than a half size of the version space, and it is more efficient than both single-hypothesis-based and committee-based active learning [19].

2. Materials and Methods

2.1. Proposed Algorithm. The flowchart of our proposed algorithm is illustrated in Figure 1. Our algorithm uses different high-level semantics as its views to generate the corresponding hypotheses. First, object detectors are trained to achieve the responses of different object classes in image regions. Furthermore, we design a cascaded online LDA (CO-LDA) as a secondary view for achieving more accurate image representation. Finally, a fine-tuned MVAL algorithm is utilized with both two high-level image semantics as its views for classifying the scene of an image.

2.2. Object Semantic-Based Image Representation. Our object semantic-based image representation is illustrated in Figure 2.

First, multiple object objectors are used to achieve the local object response maps. Second, these maps are decomposed into three spatial pyramid levels, and the maximal object responses are computed in image blocks in each spatial level, which is annotated as red blocks in Figure 2. Finally, an object response histogram is computed, which can effectively reduce the influence of object response error in the whole image. For generating the object response, a latent SVM-based detector [7] is applied for recognizing the object classes with bulk type such as car and pedestrian. Another geometric context-based detector [6] is utilized for recognizing the object classes with different textures such as tree, sky, and building.

2.3. Theme Semantic-Based Image Representation. For satisfying the dynamic update of an active learning training set, an online LDA model [20] based on stochastic gradient descent strategy is used. It adds new samples sequentially, and old samples have been no longer stored, which can achieve efficient and accurate parameter estimation in large-scale data training.

Online LDA computes the posterior probability distribution $p(\theta, z, w, \beta \mid \alpha, \eta)$ of the hidden nodes based on observed samples. It actually uses variational inference to estimate the maximum likelihood of $p(w \mid \alpha, \eta)$ based on α and η. Three variational parameters ϕ, γ, and λ follow the distributions: $\phi \sim \text{multinomial}(\varepsilon)$, $\gamma \sim \text{Dirichlet}(\varepsilon)$, and $\lambda \sim \text{Dirichlet}(\varepsilon)$. The variational distribution follows

$$q\left(\beta_{1:k}, \theta_{1:M}, z_{1:M} \mid \lambda, \gamma, \phi\right) = \prod_{i=1}^{k} \text{Dir}\left(\beta_i \mid \lambda_i\right) \cdot \prod_{d=1}^{M} q_d\left(\theta_d, z_d \mid \phi_d, \gamma_d\right). \tag{1}$$

The optimal (γ, ϕ) is solved by maximizing the lower bound in the following equation:

$$\log p(w \mid \alpha, \eta) \geq L(w, \phi, \gamma, \lambda) = E_q[\log p(\theta, z, w, \beta \mid \alpha, \eta)] - E_q[\log q(\theta, z, \beta)], \tag{2}$$

where E_q denotes the conditional mathematical expectation. Maximizing the lower bound $L(w, \phi, \gamma, \lambda)$ is equivalent to minimizing KL divergence of $q(\theta, z, \beta \mid \gamma, \phi)$ and $p(\theta, z, \beta \mid w, \alpha, \eta)$:

$$\log p(w \mid \alpha, \eta) = L(w, \phi, \gamma, \lambda) + KL(q(\theta, z, \beta) \| p(\theta, z, \beta \mid w, \alpha, \eta)), \tag{3}$$

where $L(w, \phi, \gamma, \lambda)$ is factorized as follows:

$$L(w, \phi, \gamma, \lambda) = \sum_d \left\{ E_q\left[\log p\left(w_d \mid \theta_d, z_d, \beta\right)\right] + E_q\left[\log p\left(z_d \mid \theta_d\right)\right] - E_q[\log q(z_d)] + E_q\left[\log p\left(\theta_d \mid \alpha\right)\right] - E_q[\log q(\theta_d)] + \frac{\left(E_q[\log p(\beta \mid \eta)] - E_q[\log q(\beta)]\right)}{M} \right\}. \tag{4}$$

Equation (4) can be transformed into formula (5). In equation (5), n_{dw} denotes the frequency that word w occurs in text d. $l(n_d, \phi_d, \gamma_d, \lambda)$ reflects the contribution of d for the lower bound, which is iteratively optimized by a coordinate ascent algorithm:

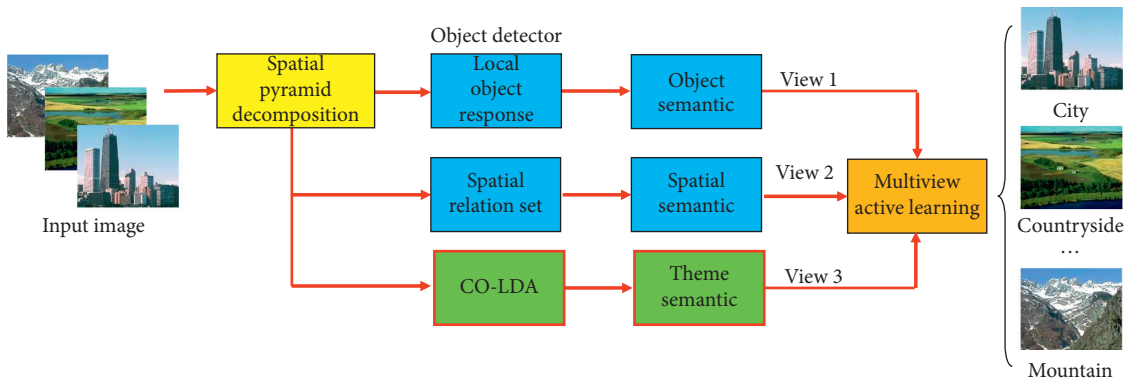

Figure 1: The flowchart of our scene classification algorithm.

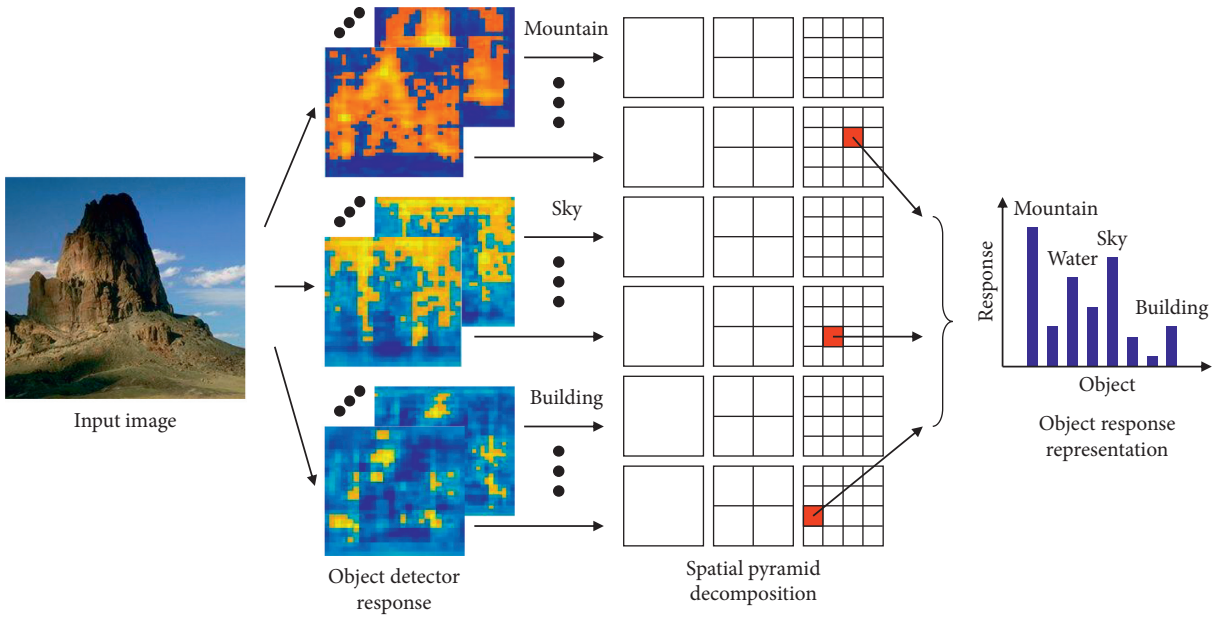

Figure 2: The flowchart of object semantic-based image representation.

$$
\begin{aligned}
L(w,\phi,\gamma,\lambda) = {} & \sum_{d}\sum_{w} n_{dw} \sum_{k} \phi_{dwk} E_q[\log \theta_{dk}] + E_q[\log \beta_{kw}] - \log \phi_{dwk} \\
& - \log \Gamma\left(\sum_{k} \gamma_{dk}\right) + \sum_{k} (\alpha - \gamma_{dk}) E_q[\log \theta_{dk}] + \log \Gamma(\gamma_{dk}) \\
& + \frac{\left(\sum_k - \log \Gamma\left(\sum_k \lambda_{kw}\right) + \sum_w (\eta - \lambda_{kw}) E_q[\log \beta_{kw}] + \log \Gamma(\lambda_{kw})\right)}{M} \\
& + \log \Gamma(K\alpha) - K \log \Gamma(\alpha) + \frac{(\log \Gamma(N\eta) - W \log \Gamma(\eta))}{M}, \\
= {} & \sum_{d} l(n_d, \phi_d, \gamma_d, \lambda).
\end{aligned}
\tag{5}
$$

ϕ_{dwk}in equation (5) is iteratively solved:

$$\begin{aligned}\phi_{dwk} &\propto \exp\left(E_q[\log(\theta_{dk})] + E_q[\log(\beta_{kw})]\right),\\ E_q[\log(\theta_{dk})] &= \Psi(\gamma_{dk}) - \Psi\left(\sum_{i=1}^{K}\gamma_{di}\right),\\ E_q[\log(\beta_{kw})] &= \Psi(\lambda_{kw}) - \Psi\left(\sum_{i=1}^{N}\lambda_{ki}\right),\end{aligned} \tag{6}$$

where digamma functionΨ is the first-order derivative of function Γ. γ_{dk} and λ_{kw} are iteratively solved in the following way: $\gamma_{dk} = \alpha + \sum_w n_{dw}\phi_{dwk}$,$\lambda_{kw} = \eta + \sum_d n_{dw}\phi_{dwk}$.

Whent^{th} vector of word frequency n_t is observed, we keep λ unchanged and update the local optimal solution of γ_t and ϕ_t in E step. In M step, ϕ_t and λ from last iteration are both used to update λ:

$$\begin{aligned}\lambda &= (1-\rho_t)\lambda + \rho_t\widetilde{\lambda},\\ \rho_t &= (\tau + t)^{-k}.\end{aligned} \tag{7}$$

$\widetilde{\lambda}$ in formula (7) is solved as follows:

$$\widetilde{\lambda}_{kw} = \eta + \frac{M}{S}\sum_s n_{\text{ts}k}\phi_{\text{ts}kw}, \tag{8}$$

where n_{ts} is s^{th} text in each batch text set, M is the number of the training text set, and $S$is the size of each batch text set. Hyperparameters α and η are updated by the Newton–Rapson method: $\alpha \longleftarrow \alpha - \rho_t\widetilde{\alpha}(\gamma_t)$ and $\eta \longleftarrow \eta - \rho_t$ $\widetilde{\eta}(\lambda)$. Here, $\widetilde{\alpha}(\gamma_t)$ is the product of Hessian matrix and gradient $\nabla_\alpha l$ of the objective function $l(n_d, \phi_d, \gamma_d, \lambda)$. $\widetilde{\eta}(\lambda)$ is the product of Hessian matrix and gradient $\nabla_\eta L$ of the objective function $L(w, \phi, \gamma, \lambda)$.

Based on online LDA, we proposed the CO-LDA theme model, which is similar with the classic SP-pLSA model in structure for enhancing the spatial correlation between different image regions. The framework of CO-LDA is illustrated in Figure 3. The main difference between CO-LDA and SP-pLSA is that different online LDAs (LDA1, LDA2, and LDA3) are applied in different spatial levels to jointly mine the theme of an image. The main advantage of CO-LDA is that it integrates the spatial correlation of objects in different image resolutions, which further improves the holistic scene understanding. The visual histogram computation in online LDA is the same as the way of object response histogram in Section 2.2, and the theme feature of each spatial block is represented by variational parameter γ of the online LDA model.

Finally, the theme feature γ of the whole image is achieved by concatenating the theme features of different blocks of different spatial pyramid levels:

$$\gamma = \left(w_1 \cdot \gamma_{L_0}\right) \oplus \left(w_2 \cdot \gamma_{L_1}\right) \oplus \left(w_3 \cdot \gamma_{L_2}\right), \tag{9}$$

where γ_{L_i} denotes the theme feature of the corresponding block in L_i^{th} pyramid level, $\oplus$ denotes the linear concatenation between feature vectors, and the weights of different spatial levels are configured as follows: $w_1 = (1/2)$, $w_2 = (1/2)$, and $w_3 = (1/4)$.

2.4. Multiview Active Learning. The MVAL referred in this paper is our previous work [21], which has two improvements in both hypothesis generation and selective sampling. First, boosting-like technique is integrated into MVAL, which uses a similar way of iterative weak classifier optimization, and the current hypothesis is boosted by weighted voting of all the hypotheses from the past queries. Furthermore, an adaptive hierarchical competition sampling is presented. In this sampling strategy, if the number of the contention samples is large, an unsupervised spectral clustering is activated to obtain the coarse spatial distribution of these contention samples in the high-dimensional feature space, and then, a multiview-based batch mode selective sampling is run based on two measures: sample uncertainty and redundancy by solving quadratic programming to determine the queried samples in each cluster.

2.4.1. Hypothesis Generation. If an active learning can select enough number of contention samples, which could improve the hypothesis in each query, the number of unlabeled samples, which are incorrectly classified, will decrease. It is quite similar with boosting technique in weak classifier optimization. The MVAL incorporates the AdaBoost algorithm into our framework to boost the generated hypothesis in each query, and the main flowchart is described in Figure 4.

In Figure 4, a support vector machine (SVM) is used as a base classifier to construct a multiview classifier, which replaces the single-view classifier in AdaBoost, and this multiview classifier in each query can be considered as a weak classifier in each iteration in AdaBoost. The hypothesis of multiview classifier $h^i(x)$ is computed by weighted voting of n SVM base classifiers $v_1, v_2, \ldots, v_n$ whose weights are $\omega_1, \omega_2, \ldots, \omega_n$. Unlike traditional query by boosting, we update the weight of each base classifier in each query and obtain the boosted hypothesis $H^i(x)$ by weighting all the hypotheses from the past queries and not from the current query only.

The detailed process of the MAVL's hypothesis generation based on AdaBoost is as follows:

(a) In iteration $h^t(x_j) = \sum_{f\in\{f_1^t, f_2^t, \ldots f_n^t\}}\omega_i^t f_i^t(x_j)$, weighted voting is used to generate the initial multiview-based hypothesis:

$$h^t\left(x_j\right) = \sum_{f\in\left\{f_1^t, f_2^t, \ldots, f_n^t\right\}}\omega_i^t f_i^t\left(x_j\right), \tag{10}$$

where $f_i^t(x_j)$ is the classification confidence of sample x_j by view i, and ω_i^t denotes the contribution of view i for classification which is determined by the soft classification error rate ε_i^t, which defines how correctly a sample is classified:

$$\varepsilon_i^t = \frac{1}{\left(\sum_{x\in L, y=1} f_i^t(x) - \sum_{x\in L, y=-1} f_i^t(x)\right)}, \tag{11}$$

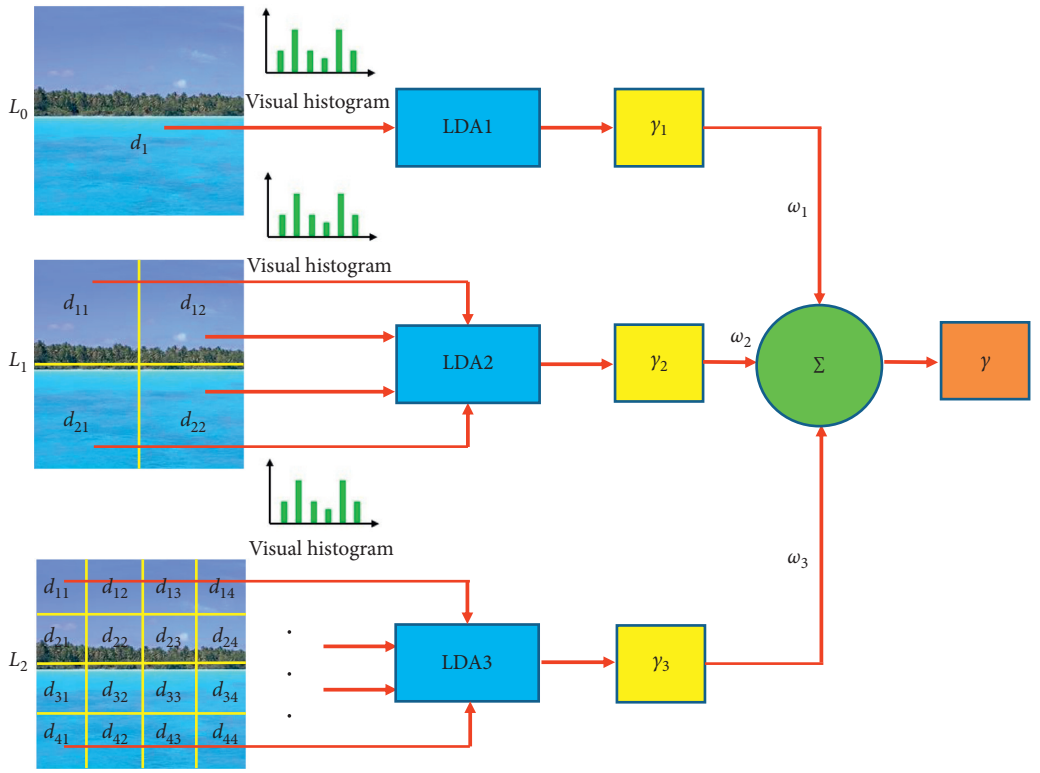

Figure 3: The framework of the CO-LDA theme model.

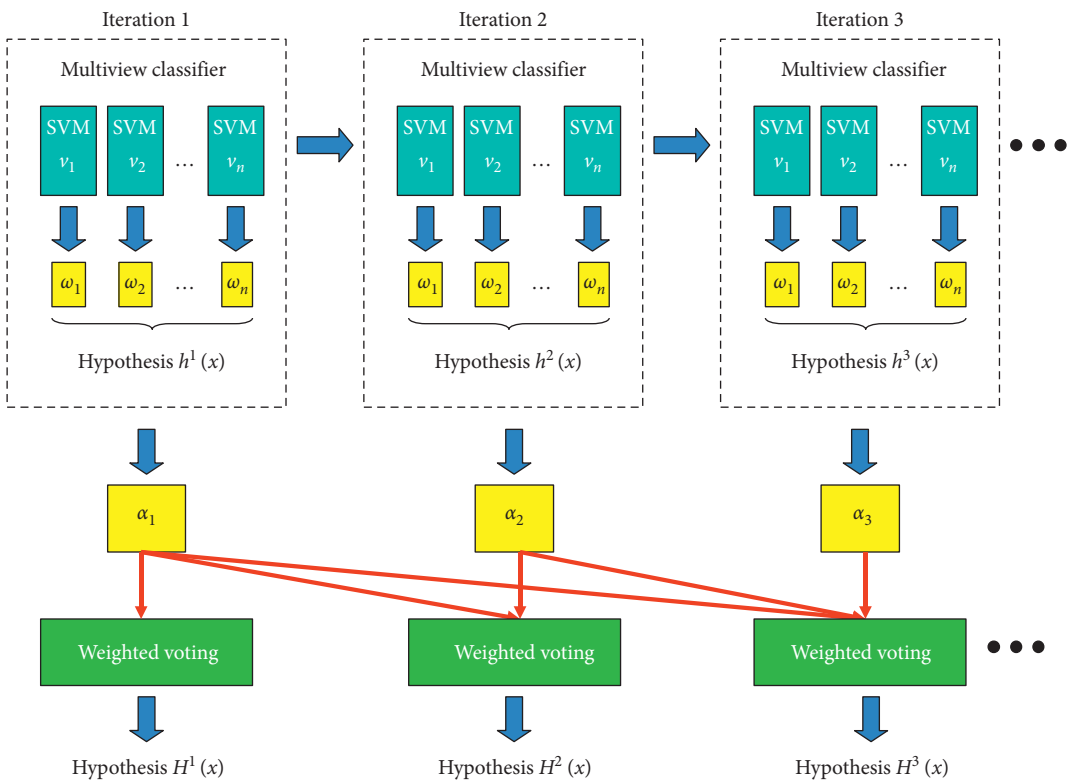

Figure 4: The framework of hypothesis generation in our MAVL [21].

where $\sum_{x\in L, y=1} f_i^t(x)$ and $\sum_{x\in L, y=-1} f_i^t(x)$ denote the sum of classification confidence of unlabeled samples, which are labeled as $y = 1$ and $y = -1$, respectively. For a "positive/negative" sample, the distance of it to the decision boundary in the "positive/negative" side reflects the degree of how correctly it is classified, and this information is utilized to calculate the error degree ε_i^t here instead of the traditional classification error calculated by the decision hypothesis in AdaBoost. Also, ω_i^t is updated through the following way: $\omega_i^t = (1/Z_1^t)\ln(1 - \varepsilon_i^t/\varepsilon_i^t)$, where Z_1^t is the normalized weight. Then, the classification confidence δ^t of the multiview classifier can be computed by the following equation:

$$\delta^t = \sum_{j=1}^{N} \beta_j^t \left| h^t(x_j) - y_j \right|. \tag{12}$$

(b) After iteration t, the size of the labeled sample set is increased as follows: $J^t = J^{t-1} \cup L^t$. J^t denotes the labeled sample set in iteration t, and L^t denotes the newly added samples after query. As we know, the size of the labeled samples set $|J^t|$ is increased during iteration in active learning. Thus, if the size of the initial labeled training set is small, the influence of $|J^t|$ should be considered when updating the weight η_t of the multiview classifier, which is illustrated by the following equation:

$$\eta_t = \frac{1}{Z_2^t}\left(\ln\left(\frac{1-\delta^t}{\delta^t}\right) + \lambda\left|J^t\right|\right). \tag{13}$$

Then, the weight of each sample is updated through the following way: $\omega_j^{t+1} = \omega_j^t \beta_t^{1-e^t}$, where $\beta_t = (\delta_t/(1-\delta_t))$, if x_j is correctly classified, $e_j = 0$, otherwise, $e_j = 1$.

(c) The final boosted hypothesis $H^t(x)$ of the queried sample x_i is equivalent to the weighted sum of all the hypotheses from the past K queries, which is defined by

$$H^t(x) = \begin{cases} 1, & \sum_{t=1}^{K} \eta_t h^t(x) \geq \frac{1}{2}\sum_{t=1}^{K} \eta_t, \\ 0, & \text{else.} \end{cases} \tag{14}$$

2.4.2. Sampling Strategy. The MVAL uses a new hierarchical competition-based sampling strategy in order to query the contention samples with high probability in different sample distributions, which is illustrated in Figure 5.

(1) *Intercluster Sampling Competition.* In the MVAL, a fast approximate spectral clustering algorithm is designed to reduce the computational complexity significantly to $O(\mathrm{KNT}) + O(K^3)$, where T is the iteration number of K mean clustering, and N is the total number of contention samples. The detailed process is illustrated as follows: (a) perform traditional K mean clustering on the contention unlabeled samples $x_1, x_2, \ldots, x_N$, compute the centroid of each cluster $y_1, y_2, \ldots, y_K$ as K representative points, and build a correspondence table to associate each x_i with the nearest cluster centroid $y_i y_i$; (b) run the normalized cut algorithm on $y_1, y_2, \ldots, y_K$ to obtain a m-way cluster membership for each of y_i; and (c) recover the cluster membership for each x_i by looking up the cluster membership of the corresponding centroid y_i in the corresponding table.

After fast spectral clustering, two intercluster sampling measures are defined: the number of samples in the cluster and its information entropy. Both measures are weighted to obtain the number of selected samples N_C^S in cluster C in the following equation:

$$N_C^S = \frac{N_T}{Z}\left[\gamma \mathrm{Num}(C) + (1-\gamma)\mathrm{Ent}(C)\right], \quad C = 1, 2, \ldots, K, \tag{15}$$

where $\mathrm{Num}(C)$ is proportional to the total number of samples N_C in cluster C, and computing $\mathrm{Ent}(C)$ is equivalent to kernel density estimation of x in cluster C. Weight $\gamma = 0.5$ reflects the impact of both measures in intercluster sampling competition, Z is the normalized factor, N_T is the total number of selected samples in the current query, and $[\cdot]$ is rounding operation.

(2) *Intracluster Sampling Competition.* In the MVAL, an efficient quadratic programming based-method [22] is utilized, which dynamically estimates the weights of the redundancy and uncertainty of an unlabeled sample in each query. It is used for intracluster selective sampling and solved by minimizing the following object function:

$$\begin{aligned} &\min_{p\in R^{n-l}} p^T \widetilde{f}_v + \frac{1}{2} p^T K_{u,u} p \\ &\text{s.t.} \quad p^T u = k,\ 0 \leq p \leq 1. \end{aligned} \tag{16}$$

Equation (16) aims to estimate the normalized parameter $p_i \in [0, 1]$, which reflects how probable the unlabeled sample is selected. $\widetilde{f}_v = (|f_v(x_1)|, \ldots, |f_v(x_l)|)^T$ is the classification confidence of sample x in v^{th} view. $x_1, \ldots, x_l$ are the queried unlabeled samples, u is a unit vector, and $k = N_C^S$ is the number of unlabeled samples in batch mode. The first part denotes the sample uncertainty in v^{th} view, and the sampling strategy tends to select the contention sample near the classification hyperplane of v^{th} view by minimizing $p^T \widetilde{f}_v$. The second part denotes the sample redundancy in v^{th} view, and the similar samples are selected by minimizing $p^T K_{u,u} p$. The sampling probability p is calculated by a convex quadratic programming, and finally, $\{p_1^v, p_2^v, \ldots, p_l^v\}$, which corresponds to $x_1, \ldots, x_l$ in v^{th} view, is obtained. For selective sampling in each cluster, the conservative sampling strategy is utilized in a classic co-testing algorithm [23].

3. Results and Discussion

In our experiment, two classic image sets (OT image set from MIT [9] and UIUC sports event image set from UIUC [24])

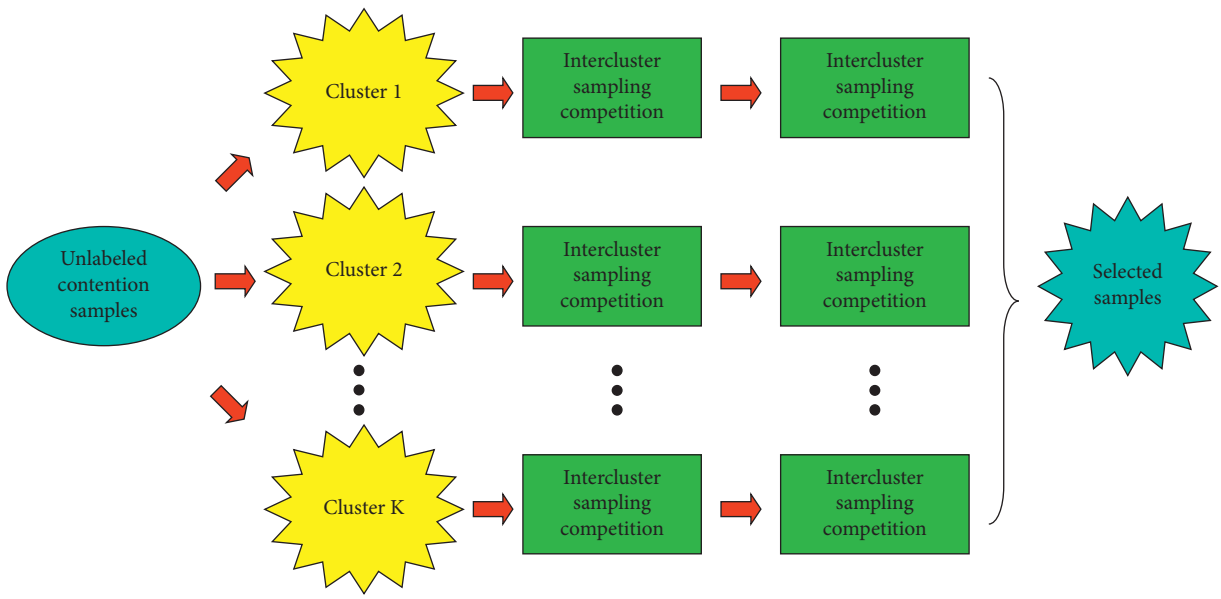

Figure 5: The framework of sampling strategy in our MAVL [21].

are used for algorithm comparison. Average classification precision (ACP) and mean of average classification precision (MACP) are both used for evaluating the performance of both CO-LDA models and multiview active learning algorithms.

3.1. Evaluation of Theme Semantic. The first experiment is designed for evaluating the performance of our proposed theme semantic. In OT and UIUC Sports datasets, the parameter configuration of the CO-LDA model is as follows: (1) $k_{\mathrm{OT}} = 0.5$, $\tau_{\mathrm{OT}} = 256$ and $k_{\mathrm{UIUC}} = 0.8$, $\tau_{\mathrm{UIUC}} = 1024$ in formula (7). (2) The batch sizes of sampled images in MVAL are $S_{\mathrm{OT}} = S_{\mathrm{UIUC}} = 512$.

We observe MACP variation of the CO-LDA model by changing the numbers of both theme and visual word: $T = 20, 30, 40, 50$ and $W = 200, 500, 800, 1200, 1500$, and a total of twenty groups of (T, W) are obtained. In Figure 6, we find that (T, W) curves for both datasets show the similar trends that MASP increase first and then decrease. Thus, in our CO-LDA model, we set $T_{\mathrm{OT}} = 30, W_{\mathrm{OT}} = 500$ and $T_{\mathrm{UIUC}} = 40, W_{\mathrm{UIUC}} = 1200$.

Figures 7(a)-7(b) and 8(a)-8(b) show the probability distributions of different themes by CO-LDA in OT and UIUC Sports image datasets.

In the OT image set, we can see that there are significant differences between four scene classes "Highway," "Forest," "Mountain," and "Tall building" in theme probability distributions, and multiview SVM classifier works well in scene classification. In the UIUC Sports image set, the theme probability distributions are very similar in four scene classes "Bocce," "Croquet," "Polo," and "Snowboarding," which significantly increases the difficulty of scene classification.

Furthermore, we compare the CO-LDA model with traditional LDA [9] and SP-pLSA [8] models, and the performance comparison of three theme models is shown in

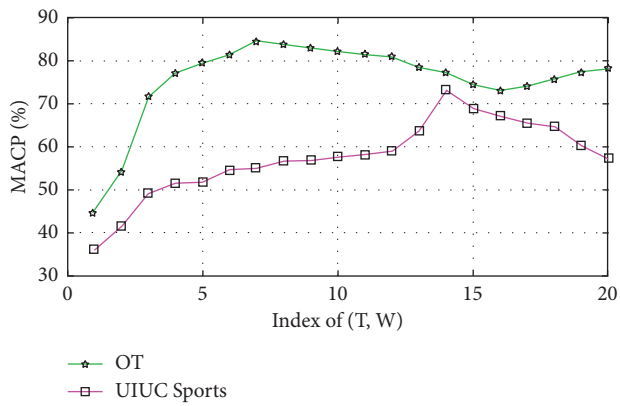

Figure 6: MACP curves with different groups of the theme and visual word in OT and UIUC Sports datasets.

Table 1. N1 ~ N8 denote the following eight natural scene classes: "Coast," "Forest," "Mountain," "Open country," "Highway," "Inside city," "Tall building," and "Street." S1 ~ S8 denote the following eight event scene classes: "Badminton," "Bocce," "Croquet," "Polo," "Rock Climbing," "Rowing," "Sailing," and "Snowboarding."

In the LDA model, each image is divided into 11×11 blocks, and 5 pixels are overlapped between neighbored blocks. For feature representation, gray-scale SIFT descriptors are sparsely sampled, and means of three color channels are calculated. The numbers of the theme and visual word are $T_{\mathrm{OT}} = 30, W_{\mathrm{OT}} = 200$ and $T_{\mathrm{OT}} = 50, W_{\mathrm{OT}} = 800$ by cross validation. In the SP-pLSA model, the ways of image division and feature representation are the same as the LDA model. The numbers of the theme and visual word are $T_{\mathrm{OT}} = 25, W_{\mathrm{OT}} = 1200$ and $T_{\mathrm{OT}} = 50, W_{\mathrm{OT}} = 1500$ by cross validation.

In the OT image set, CO-LDA achieves both higher ACP and MACP than SP-pLSA in six scene classes except "Mountain" and "Inside city." LDA performs the worst in all

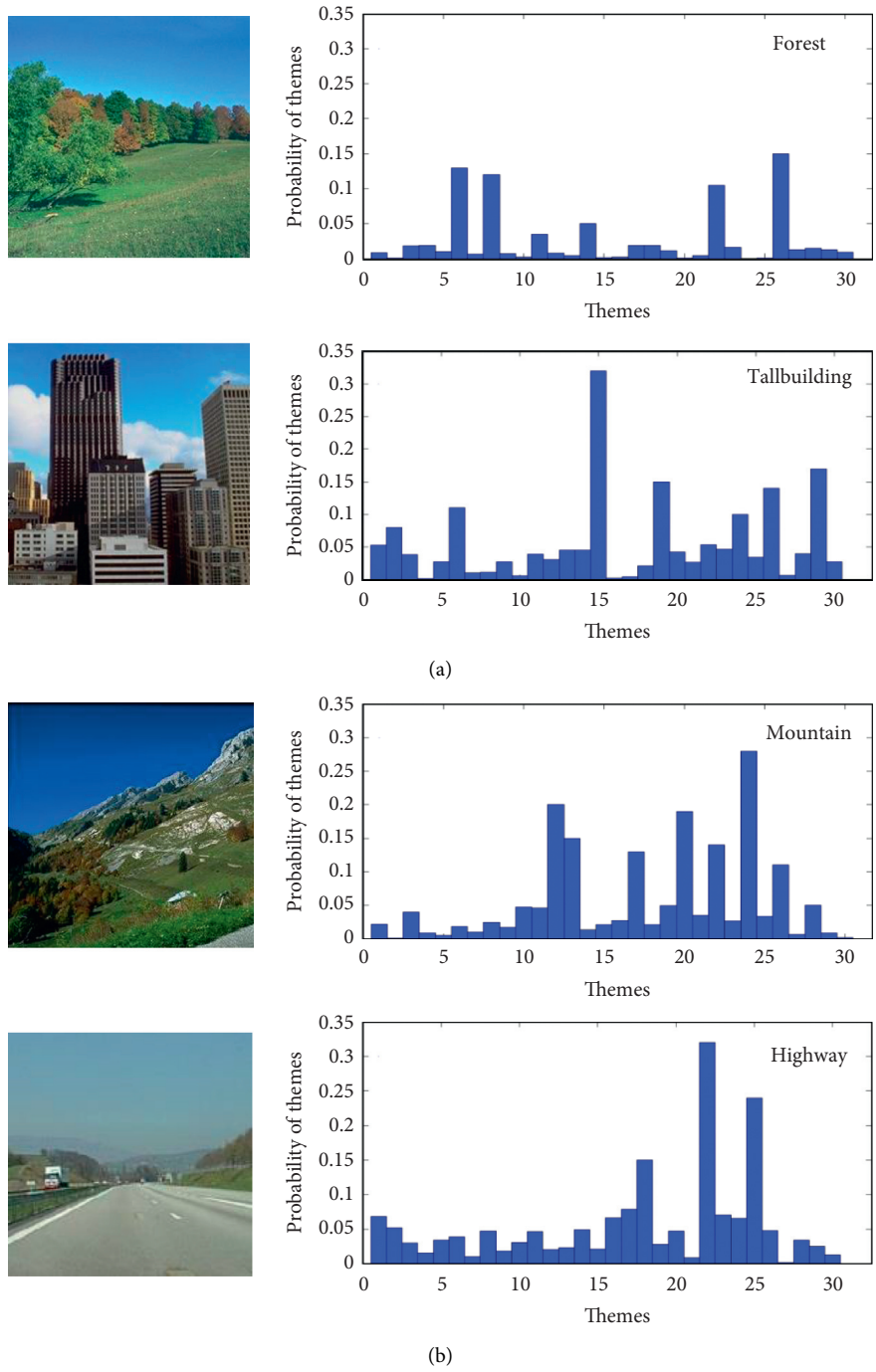

Figure 7: The theme probability distribution in different scene classes (OT image dataset).

of scene classes except "Street." It is easy to conclude that CO-LDA can achieve more accurate scene semantics than other two classic methods. In the UIUC Sports image set, CO-LDA achieves the highest ACP in the following three event classes: "Croquet," "Polo," and "Rowing," and SP-pLSA achieves the highest ACP in the following three event classes: "Bocce," "Rock Climbing," and "Snowboarding." But in the event classes "Badminton" and "Sailing," in which LDA has the highest ACP, CO-LDA still performs better than SP-pSLA. Thus, we can conclude that our proposed CO-LDA also have slightly better performance in theme mining than the two classic image representation methods.

3.2. Evaluation of MVAL. In the second experiment, we compare our algorithm with other single-view active learning algorithm with both high-level semantics and low-level features for scene classification. In our initial labeled training set, label size = 150, batch size = 20, and iteration = 10.

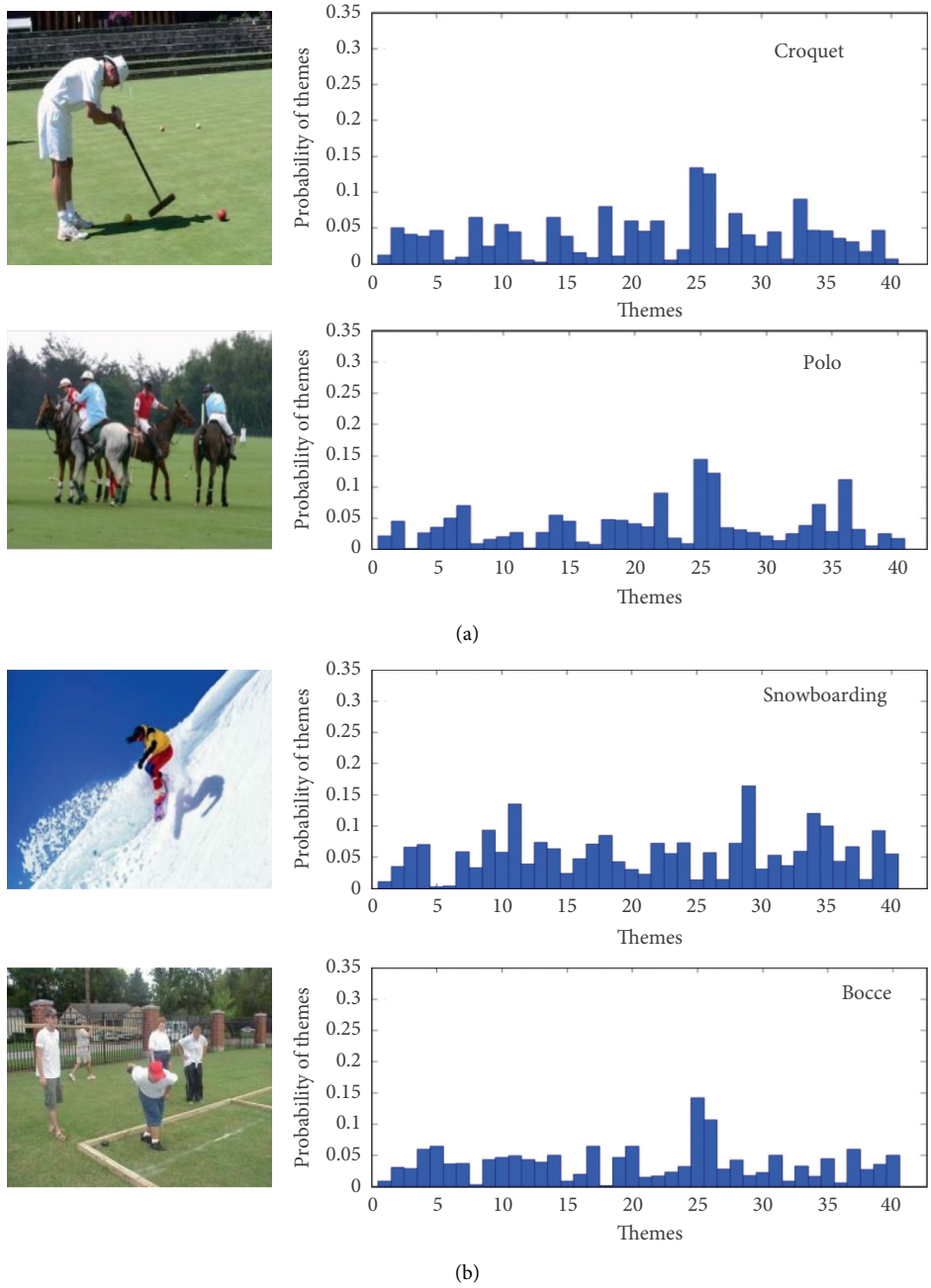

Figure 8: The theme probability distribution in different scene classes (UIUC image dataset).

Our proposed algorithm MVAL^{HS} (MVAL and HS denote MVAL [21] and two proposed high-level semantics, respectively) is compared with the following four algorithms: (1) MVAL^{LS} (LS denotes low-level image features): in MVAL, both means of three color channels and densely sampled color-SIFT descriptors are concatenated as a feature vector for image representation. (2) AL^{QP} [22]: a single-view SVM active learning by QP-based selective sampling, which relies on the sample uncertainty and redundancy. (3) Diff^{WS} [25–30]: a disagreement-based active learning from weak and strong labelers. (4) Graph^{GP} [23]: a graphical model-based active learning with robust Gaussian process. The feature representations of AL^{QP}, Diff^{WS}, and Graph^{GP} are the same as MVAL^{LS}. The performance comparison of the five active learning algorithms is shown in Table 2.

From Table 2, it is easily found that our algorithm MVAL^{HS} has the highest MACP in almost all scene classes than the other four algorithms in both image sets, which demonstrates that high-level semantics can achieve more significant improvement in holistic scene understanding than

Table 1: Performance comparison of different theme models.

ACP	N1	N2	N3	N4	N5	N6	N7	N8	*MACP*
(a) OT image dataset									
LDA	73.24 ± 1.16	86.69 ± 0.55	74.83 ± 0.72	64.42 ± 1.47	68.13 ± 1.85	80.75 ± 1.60	82.10 ± 0.50	84.74 ± 1.31	76.56 ± 1.15
SP-pLSA	82.40 ± 0.58	91.59 ± 0.34	87.74 ± 1.79	73.05 ± 1.04	79.66 ± 0.98	89.27 ± 0.59	89.68 ± 1.42	79.06 ± 1.55	83.80 ± 1.04
CO-LDA	86.32 ± 1.39	94.22 ± 1.53	85.17 ± 0.21	73.70 ± 0.30	83.53 ± 0.52	87.39 ± 1.26	90.74 ± 1.38	82.47 ± 0.41	86.26 ± 0.88
ACP	S1	S2	S3	S4	S5	S6	S7	S8	*MACP*
(b) UIUC Sports image dataset									
LDA	79.84 ± 2.28	63.86 ± 0.94	59.74 ± 1.53	65.49 ± 0.59	73.01 ± 0.29	61.65 ± 0.98	84.92 ± 1.38	58.22 ± 0.03	68.30 ± 1.00
SP-pLSA	75.25 ± 1.44	69.14 ± 0.51	67.19 ± 1.11	62.70 ± 1.73	86.32 ± 1.50	72.54 ± 0.13	80.75 ± 0.76	75.34 ± 1.02	73.64 ± 1.13
CO-LDA	77.62 ± 0.16	67.48 ± 0.74	68.73 ± 0.86	71.48 ± 2.13	82.75 ± 1.04	77.81 ± 0.07	83.43 ± 0.42	72.89 ± 0.55	75.27 ± 0.75

Note. Bold values represent the best performance of the algorithms corresponding to each class.

Table 2: Performance comparison of different active learning algorithms.

DS	Algorithm	N1	N2	N3	N4	N5	N6	N7	N8
OT image dataset									
	AL^{QP}	70.52 ± 1.80	85.65 ± 1.50	74.94 ± 1.33	53.52 ± 2.14	77.34 ± 074	80.04 ± 017	79.17 ± 1.34	75.00 ± 0.85
	$Diff^{WS}$	72.43 ± 1.67	84.19 ± 1.59	76.53 ± 1.73	55.14 ± 0.74	72.88 ± 0.98	78.58 ± 1.12	81.40 ± 0.88	74.30 ± 1.00
	$Graph^{GP}$	71.52 ± 1.44	86.32 ± 0.99	74.30 ± 0.15	52.05 ± 1.54	75.85 ± 1.17	79.53 ± 2.03	78.76 ± 1.89	76.75 ± 0.44
	$MVAL^{LS}$	73.23 ± 0.55	85.57 ± 1.76	77.53 ± 1.59	53.85 ± 0.95	77.34 ± 1.22	81.40 ± 2.00	81.87 ± 1.36	77.52 ± 0.53
	$MVAL^{HS}$	77.25 ± 0.88	90.41 ± 1.01	80.37 ± 1.55	60.49 ± 0.29	82.14 ± 0.72	87.38 ± 1.13	85.91 ± 0.14	81.73 ± 0.66
DS	Algorithm	S1	S2	S3	S4	S5	S6	S7	S8
UIUC-Sports image dataset									
	AL^{QP}	68.54 ± 1.41	49.42 ± 0.32	45.90 ± 1.40	55.17 ± 0.62	67.19 ± 0.88	60.20 ± 1.52	75.19 ± 1.07	51.09 ± 1.21
	$Diff^{WS}$	72.22 ± 1.50	50.98 ± 1.22	48.79 ± 0.59	57.33 ± 1.31	67.00 ± 0.55	57.53 ± 0.78	71.01 ± 0.16	54.58 ± 1.50
	$Graph^{GP}$	69.18 ± 1.37	54.10 ± 0.76	47.52 ± 2.13	53.75 ± 0.89	69.04 ± 0.34	62.55 ± 1.04	74.10 ± 1.00	55.22 ± 0.28
	$MVAL^{LS}$	71.18 ± 1.39	53.88 ± 0.55	49.62 ± 0.54	57.93 ± 1.65	71.05 ± 1.75	62.98 ± 1.70	75.59 ± 1.04	57.18 ± 1.24
	$MVAL^{HS}$	76.75 ± 1.39	55.44 ± 1.60	53.96 ± 1.23	59.78 ± 0.77	70.66 ± 0.52	66.40 ± 1.08	78.05 ± 1.51	60.68 ± 0.41

Note. Bold values represent the best performance of the algorithms corresponding to each class.

traditional low-level image features. Furthermore, we can see that $MVAL^{LS}$ performs better in most cases than other three single-view algorithms, which also means that multiple view setting can successfully result in larger decrease in the size of the version space than traditional single-view active learnings due to its independent and diverse views.

4. Conclusion

This paper proposed a MVAL-based scene classification algorithm, which applies two different high-level image semantics to generate the corresponding hypotheses. Different object detectors are first trained to achieve the responses of different object classes as object semantic. Furthermore, a CO-LDA model is proposed for achieving more accurate theme semantic by integrating the spatial correlation of objects in different image resolutions, which improves the holistic scene understanding. With the help of both two independent views, our MVAL algorithm has potential to not only handle large-scale data training but also improve the performance of scene classification.

Data Availability

All data utilized in our research can be accessed from the following website: https://archive.ics.uci.edu/ml/datasets/Corel+Image+Features and http://vision.stanford.edu/lijiali/event_dataset/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was funded by the Natural Science Foundation of China (Grant nos. 41571299 and 11601339), Key Research and Development Plan of Zhejiang Province (Grant no. 2018C01086), Open Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China (no. ICT20047), Zhejiang Provincial Natural Science Foundation of China (Grant no. LY18F020025), and National Thousand Talents Program (Grant no. Y474161).

References

[1] P. F. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 32, no. 9, pp. 145–175, 2001.

[2] A. Oliva and A. Torralba, "Chapter 2 Building the gist of a scene: the role of global image features in recognition," *Progress in Brain Research*, vol. 30, no. 4, pp. 23–36, 2006.

[3] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[4] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the conference of computer vision and pattern recognition*, Honolulu, HI, USA, 2017.

[5] H. Guo, K. Zhang, X. C. Fan, H. K. Yu, and S. Wang, "Visual attention consistency under image transforms for multi-label image classification," in *Proceedings of the conference of computer vision and pattern recognition*, Long Beach, CA, USA, 2019.

[6] D. Hoiem, A. N. Stein, A. A. Efros, and M. Hebert, "Recovering occlusion boundaries from a single image," in *IEEE 11th International conference on computer vision*, pp. 1–8, Rio de Janeiro, Brazil, 2007.

[7] P. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1627–1645, 2010.

[8] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 712–727, 2008.

[9] L. Fei-Fei and P. Peronma, "A bayesian hierarchical model for learning natural scene categories," in *IEEE computer society conference on computer vision and pattern recognition*, pp. 524–531, San Diego, CA, USA, 2005.

[10] Y. Tokozume, Y. Ushiku, and T. Harada, "Between-class learning for image classification," in *Proceedings of the conference of computer vision and pattern recognition*, Salt Lake City, UT, USA, 2018.

[11] Z. M. Chen, X. S. Wei, P. Wang, and Y. W. Guo, "Multi-label image recognition with graph convolutional networks," in *Proceedings of the conference of computer vision and pattern recognition*, Long Beach, CA, USA, 2019.

[12] X. Y. Zhang, S. H. Du, and Y. Zhang, "Semantic and spatial co-occurrence analysis on object pairs for urban scene classification," *IEEE Journal of Selected Ttopics in Applied Earth Observations and Remote Sensing*, vol. 11, pp. 2630–2643, 2018.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, 2015.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, "SSD: single shot multibox detector," in *European conference on computer vision*, Amsterdam, Netherlands, 2016.

[15] T. Y. Lin, P. Dollar, and R. Girshick, "Feature pyramid networks for object detection," in *Proceedings of the conference of computer vision and pattern recognition*, Honolulu, HI, USA, 2017.

[16] P. Liu, H. Zhang, and K. B. Eom, "Active deep learning for classification of hyperspectral images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 712–724, 2017.

[17] K. Wang, D. Y. Zhang, Y. Li, R. M. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[18] L. Yang, Y. Z. Zhang, J. X. Chen, S. Y. Zhang, and D. Z. Chen, "Suggestive annotation: a deep active learning framework for biomedical image segmentation," in *Proceedings of the conference of computer vision and pattern recognition*, Honolulu, HI, USA, 2017.

[19] W. H. Yang, G. Q. Liu, L. Zhang, and E. H. Chen, "Multi-view learning with batch mode active selection for image retrieval," in *Proceedings of the 21st international conference on pattern recognition*, pp. 979–982, Tsukuba, Japan, 2012.

[20] M. D. Hoffman, D. M. Blei, and F. Bach, "Online learning for latent dirichlet allocation," *Proceedings of Neural Information Processing Systems*, pp. 1–9, 2010.
[21] T. Z. Yao, P. An, and J. T. Song, "Multi-view active learning based on weighted hypothesis boosting and hierarchical competition sampling," *Acta Electronica Sinica*, vol. 45, no. 1, pp. 46–53, 2017.
[22] S. C. H. Hoi, R. Jin, J. K. Zhu, and M. R. Lyu, "Semi-supervised svm batch mode active learning for image retrieval," in *IEEE conference on computer vision and pattern recognition*, pp. 1–7, Anchorage, AK, USA, 2008.
[23] C. C. Long and G. Hua, "Multi-class multi-annotator active learning with robust Gaussian process for visual recognition," in *IEEE international conference on computer vision*, pp. 2839–2847, Santiago, Chile, 2015.
[24] L. J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *IEEE 11th International conference on computer vison*, pp. 1–8, Rio de Janeiro, Brazil, 2007.
[25] C. C. Zhang and K. Chaudhuri, "Active learning from weak and strong labelers," *Advances in Neural Information Processing Systems*, 2015.
[26] X. M. Zhang, T. T. Wang, J. Q. Qi, H. C. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proceedings of the conference of computer vision and pattern recognition*, Salt Lake City, UT, USA, 2018.
[27] T. Zhao and X. Q. Wu, "Pyramid Feature attention network for saliency detection," in *Proceedings of the conference of computer vision and pattern recognition*, Long Beach, CA, USA, 2019.
[28] H. L. Zheng, J. L. Fu, Z. J. Zha, and J. B. Luo, "Looking for the devil in the details: learning trilinear attention sampling network for fine-grained image recognition," in *Proceedings of the conference of computer vision and pattern recognition*, Long Beach, CA, USA, 2019.
[29] L. J. Li, H. Su, E. P. Xing, and L. Fei-Fei, "Object bank: a high-level image representation for scene classification & semantic feature sparsification," *Proceedings of Neural Information Processing Systems*, pp. 1–9, 2010.
[30] I. Muslea, S. Minton, and C. A. Knoblock, "Active learning with multiple views," *Journal of Artificial Intelligence Research*, vol. 27, no. 1, pp. 203–233, 2006.