21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Multi-View Learning with Batch Mode Active Selection for Image Retrieval

Wenhui Yang, Guiquan Liu, Lei Zhang and Enhong Chen
University of Science and Technology of China, Anhui, China
{*yangwh,stone*}*@mail.ustc.edu.cn,* {*gqliu,cheneh*}*@ustc.edu.cn*

Abstract

With the explosive growth of Internet image data, labeling image data for image retrieval has become an increasingly onerous task. To that end, we proposed a novel multi-view learning with batch mode active learning framework, MV-BMAL, for improving the performance of image retrieval. Specifically, color, texture and shape features are extracted and considered as uncorrelated and sufficient views of an image, then each classifier is trained on these views respectively, and the schema makes full use of the classification results of each unlabeled samples to find out the most informative and representative samples for automatically or manually labeling. Finally, we evaluate MV-BMAL on benchmark data sets, and the experimental results show that our proposed MV-BMAL algorithm significantly outperforms the previous methods.

1. Introduction

With the massive growth of the volume of digital image data, Content-Based Image Retrieval (CBIR) has become an important and challenging research topic in recent years [4]. Most of existing algorithms adopt low-level features, such as color, texture and shape, as the representation of images. However, it is hard to describe the semantic concepts with those low-level features, which means that there is a semantic gap between low-level features and the semantic interpretation [7].

In order to narrow the semantic gap and reduce the effort involved in acquiring labeled images, many relevance feedback algorithms have been proposed [2, 5, 6]. The relevance feedback algorithm is that users provide feedbacks regarding the relevance or irrelevance of the current retrieval results, then the classifiers will be trained based on the feedback results to further improve the retrieval performance. One important issue of feedback algorithms is that how to obtain the most informative and representative unlabeled images so that the retrieval performance could be improved most efficiently. To address this issue, Tong et al. [5] proposed an algorithm named SVM-AL to select the feedback images which are closest to the support vector boundary. After that, TSVM-AL proposed by Wang et al. [6] and Co-SVM proposed by Cheng et. al [2] using different strategies to optimize the SVM-AL algorithm. However, in those algorithms, feedback images are selected in a batch which contains top-k images closest to decision boundary, so that those samples could be similar or even identical to each other and can not provide additional features.

On the other hand, the feature representation of an image is usually a combination of diverse sub-features, such as color, texture and shape. Due to every sub-feature's distribution is always conditionally independent, the contribution of each sub-feature is obviously different. Thus, it is conductive to train each classifier based on each feature subspace individually and then measure the consistency of the all classification results for making final decisions[3].

To address these above issues, we proposed a novel schema MV-BMAL combining multi-view learning and batch mode active learning for image retrieval. First, ***multi-view learning*** is a classic semi-supervised mechanism which reduces the amount of labeled samples required for learning by exploring complementary information from disjoint sub-sets of features [3]. Specifically, color, texture and shape features are extracted and considered as uncorrelated and sufficient views of an image, and then each classifier is trained on those views respectively. After that, multi-view learner can make full use of those sub-features' information and utilize the agreement among different learners to obtain the most informative and representative samples. Second, ***batch mode active learning*** is a strategy to select the top-k most ambiguity samples for user manually labeling. The traditional *top-k active strategy* is simply choosing samples closest to the decision boundary in which some feedback samples could be similar or even identical to each other, so that they could not provide additional information for model updating.

978-4-9906441-0-9

Therefore, we proposed a new strategy named **Spectral Fuzzy Cuts(SF-cuts)** to get the most informative and representative feedback samples without redundancy.

Our paper is organized as follows. Section 2 presents the proposed framework. Section 3 illustrates the experimental results and the conclusions are presented in Section 4.

2. Multi-view Learning with Batch Mode Active Selection Framework

2.1. Overview of Our Proposed Framework

In our MV-BMAL model, we use a two-phase schema to select the unlabeled samples which can be automatically or manually labeled from the database, thus it can attack the problems of insufficient training data. Fig. 1 illustrates the framework of our model.

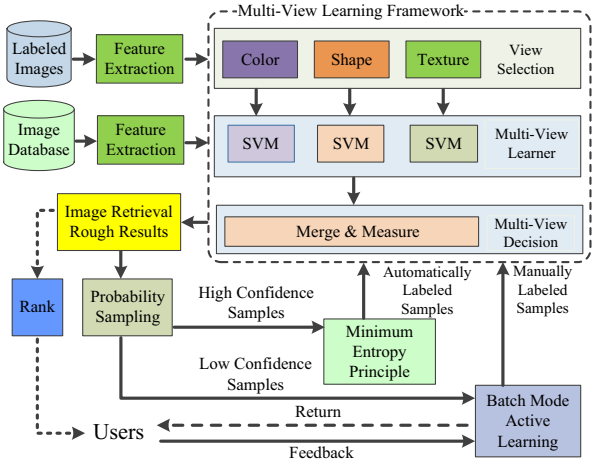


Figure 1. Multi-View Learning with Active Selection Framework

In the first phase, we extract the color, shape and texture features from the total non-independent view, then we employ each classifier (SVM) on each feature subspace to learn several rough decision boundaries based on the initial labeled samples $\mathcal{L}$. Next, we calculate the view-similarity of each unlabeled samples to generate rough sets containing the low and high confidence samples ($\mathcal{U}_{low}$ and $\mathcal{U}_{high}$) through the statistical methods (Section 2.2). In the second phase, we apply different strategies to automatically label the subset samples from $\mathcal{U}_{high}$ (Section 2.3) or manually label the subset samples from $\mathcal{U}_{low}$ (Section 2.4), and then add those labeled samples into training data for the next round execution.

2.2. Multi-View Learning

Suppose that there are l labeled data samples $\mathcal{L}$: $(x_1, y_1), \ldots, (x_l, y_l)$ and u unlabeled data samples $\mathcal{U}$: $x_{l+1}, \ldots x_{l+u}$, where $l \ll u$, and assume $x_k = (x_k^{(1)}, \ldots, x_k^{(V)})$ be a multi-view sample with V views, then we regard $f_i : x^{(i)} \rightarrow Y$ $(i = 1...V)$ as the classifier works in each view. Multi-view learning techniques train a set of classifiers $\{f_i\}$ by maximizing their consensus on the unlabeled data. Let $x_k \in U$, and $p_{kc}^{(v)}$ denotes that the probability of x_k belongs to the c-th class in view v.

$$p_{kc}^{(v)} = P(f_v(x_k^{(v)}) = c | x_k^{(v)}) \tag{1}$$

For comprehensive evaluation of the confidence of each sample's classification results over all views, we use the Gaussian distance to measure the similarity of the results. For example, suppose that $x_k \in U$ and $p_{k\cdot}^{(v)} = (p_{k1}^{(v)}, \ldots, p_{kc}^{(v)})^T$, the multi-view similarity of x_k can be expressed as follows:

$$Sim(x_k) = \sum_{\{i,j\} \subseteq V \wedge i \neq j} \exp(-\alpha \left\| p_{k\cdot}^{(i)} - p_{k\cdot}^{(j)} \right\|_2^2) \tag{2}$$

Then we can obtain μ, σ by fitting $Sim(x_k)$ $(k = 1...|U|)$ for the Normal Distribution and get low fractile and high fractile respectively (i.e. $\alpha = 0.05$):

$$\begin{aligned} \mathcal{U}_{low} &= \{x \in \mathcal{U} : Sim(x) < \mu - \Phi^{-1}(1 - \frac{\alpha}{2}) \times \sigma\} \\ \mathcal{U}_{high} &= \{x \in \mathcal{U} : Sim(x) > \mu + \Phi^{-1}(1 - \frac{\alpha}{2}) \times \sigma\} \end{aligned} \tag{3}$$

The samples of $\mathcal{U}_{high}$ with highly agreement over all views, are always away from the classification margin with minimum label ambiguity. They can be automatically labeled easily, and added to the original labeled data set $\mathcal{L}$. On the contrary, the samples of $\mathcal{U}_{low}$ with highly disagreement over all views, which cause a lot of uncertainties of the retrieval results. Thus, we can adopt active learning to eliminate the adverse effects of these samples which are closet to the decision boundary.

2.3. Minimum Entropy Principle for High Confidence Samples

We employ the minimum entropy principle to obtain the most appropriate label for the most reliable high confidence samples from the $\mathcal{U}_{high}$. Let $l_{kc}^{(v)}$ denote that the sample x_k in the v-th view is assigned to label c, when $p_{kc}^{(v)}$ has the maximum value in the $p_{k\cdot}^{(v)}$ set.

To measure the level of disagreement, we use *vote entropy* to metric the purity of the classification results over all views.

$$VE_x = -\sum_{i=1 \wedge vt(l_{xi}^{(\cdot)}) \neq 0}^{C} \frac{vt(l_{xi}^{(\cdot)})}{V} \log \frac{vt(l_{xi}^{(\cdot)})}{V} \tag{4}$$

where $vt(l_{xi}^{(\cdot)})$ indicates the number of votes about instance x is assigned to the i-th class in each view. Let

VE^* denote the subset of $\mathcal{U}_{high}$ used for automatically labeling:

$$VE^* = \{x \in \mathcal{U} : VE_x \leq g(VE, \alpha)\} \quad (5)$$

where $g(VE, \alpha)$ is a lower confidence limit indicator and $g(VE, \alpha) = E(VE) - \Phi^{-1}(1 - \frac{\alpha}{2}) \times Var(VE)$, Φ is the standard normal distribution, $E(VE)$ and $Var(VE)$ are the expectation and variance of VE respectively and α is the confidence level. If $VE_x \in VE^*$, sample x is then automatically assigned to the class label which has a maximum number of votes.

Finally, labeled samples VE^* are added to the training set. These samples together with initial labeled samples to reduce the time consuming of obtaining the additional labeled samples and build better classifiers for improving the model performance.

2.4. Batch Model Active Learning for Low Confidence Samples

We employ Batch Model Active Learning (BMAL) with ***Spectral Fuzzy Cuts (SF-cuts)*** to select the most valuable samples for manually labeling. Due to many of the images' features are often irrelevant and the same images have different topics, we first use ***Spectral method*** to project our feature space into a low-dimensional space, then apply fuzzy partitioning algorithms for graph cuts and find out the most valuable feedback samples in each subgraph.

We construct a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the vertex set $\mathcal{V} = \{x_1, \ldots, x_m\}^T$ in $\mathcal{R}^{m \times d}$ denotes the samples of $\mathcal{U}_{low}$, and edge set $\mathcal{E}$ denotes the similarity value of each sample. We apply the SF-cuts strategy to divide the entire graph $\mathcal{G}$ into different subgraphs according to the size of $\mathcal{Q}$, and then to select the most representative sample in each subgraph.

$$A_{ij} = \begin{cases} \exp(\frac{-d^2(x_i, x_j)}{\sigma_i \sigma_j}) & i \neq j \\ 0 & i = j \end{cases} \quad (6)$$

We can define a affinity matrix $A_{m \times m}$, its element A_{ij} describes the similarity value between sample x_i and x_j according to Eq. 6. Let $\sigma_i = d(x_i, x_{i_l})$, the x_{i_j} denote the l-th (ie. $l = 5$) neighbor of sample x_i, and the $d(a, b)$ is a Euclidean distance function. Meanwhile, the degree matrix D is a diagonal matrix which element is $D_{ii} = \sum_{j=1}^{m} A_{ij}$.

Then, we normalize the affinity matrix $A_{m \times m}$ using $L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ and apply the Principal Component Analysis (PCA) method to find the K largest principal components, $PC_{m \times k} = \{p_1, p_2, \ldots, p_k\}, p_i \in \mathcal{R}^m$. Next, we can form the spectral presentation matrix $Y_{m \times k}$ of $\mathcal{U}_{low}$ by re-normalizing PC with Eq. 7.

$$y_{ij} = PC_{ij} \Big/ \left(\sum_{j=1}^{m} PC_{ij}^2\right)^{\frac{1}{2}} \quad (7)$$

Suppose that $C = \{c_1, c_2, \ldots, c_s\}$ is the centroids set corresponding to the subgraph set $\tilde{G} = \{\tilde{G}_1, \ldots \tilde{G}_s\}$, $\mathcal{D}$ is the fuzzy membership matrix, d_{ij} denotes the fuzzy membership degree of sample y_j belongs to $\tilde{G}_i$, and $\mathcal{D} = \{[d_{ij}]_{s \times m} | \forall i, \forall k, 0 \leq d_{ij} \leq 1, \sum_{i=1}^{s} d_{ij} = 1, \sum_{j=1}^{m} d_{ij} > 0\}$, w is the degree of mixture. The objective is to obtain the s fuzzy subgraphs for Y by minimize the evaluation function $J_{fuz}(D, Y)$,

$$\min J_{fuz}(\mathcal{D}, Y) = \sum_{i=1}^{s} \sum_{j=1}^{m} (d_{ij})^w \|y_j - c_i\|_2^2 \quad (8)$$

Bezdek [1] applied iterative procedures to obtain an approximate solution.

$$d_{ij} = \frac{1}{\sum_{r=1}^{s} \left(\frac{\|c_i - y_j\|}{\|c_r - y_j\|}\right)^{\frac{2}{w-1}}} \quad c_i = \frac{\sum_{j=1}^{m} (d_{ij})^w y_j}{\sum_{j=1}^{m} (d_{ij})^w} \quad (9)$$

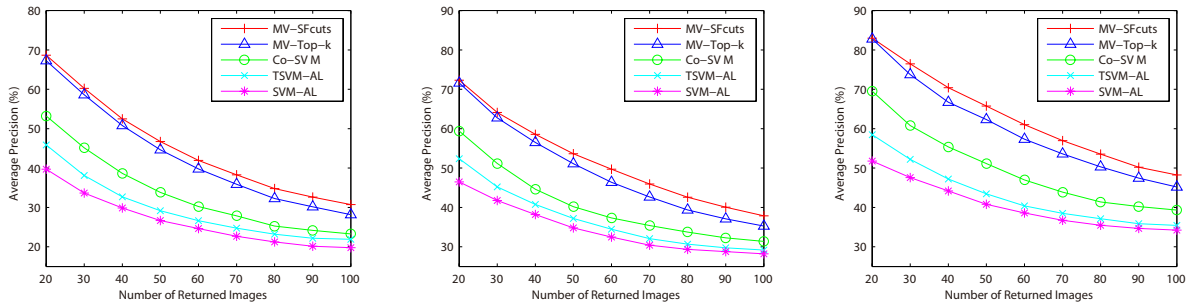
Then, we return the nearest samples from the centroids C, and ask users for manually labeling. Finally, those labeled samples would be added to the training set $\mathcal{L}$ for building better classifiers.

3. Experimental Results and Analysis

Data Set. All the experiments are performed on a real-world color images data set (50-Category) from Corel image CDs[1]. Every category consists of 100 images and has different semantic topics, such as flower, bus, beach and so on. The color features are presented by a 9-dimensional vector which includes color probability distribution, color mean, variance and skewness in each channel (HSV). The edge direction histogram which is extracted by Canny operator, denotes the shape information, and it contains two 9-dimensional vectors in horizontal and vertical directions, respectively. To describe the texture features, 9 Gabor filters with different angles are applied in gray image to obtain various feature detection images, and then we compute the entropy of each image for forming a 9-dimensional texture vector.

Performance Evaluation. For all algorithms, 10 images of each topic are randomly selected for assignment of labels as the initial training data set L (500 images in total) and 3 rounds relevance feedback are conducted for each schema. In each round, several high confidence images are labeled automatically, and 10 representative low confidence images are presented to users for manually labeling. We calculate the **Average Precision** of all query topics with different quantity of the returned images in each execution.

[1] http://www.corel.com/corel/



(a) Feedback round 1 (b) Feedback round 2 (c) Feedback round 3

Figure 2. Experimental results of the image set.

	Top 20				Top 40				Top 60				Top 80				Top 100			
	SVM AL	TSVM AL	Co SVM	MV-BMAL	SVM AL	TSVM AL	Co SVM	MV-BMAL	SVM AL	TSVM AL	Co SVM	MV-BMAL	SVM AL	TSVM AL	Co SVM	MV-BMAL	SVM AL	TSVM AL	Co SVM	MV-BMAL
Round 1	39.72	45.86	53.21	**68.67**	29.86	32.68	38.64	**52.48**	24.64	26.62	30.21	**41.94**	21.24	23.20	25.28	**34.79**	19.76	21.91	23.28	**30.71**
Round 2	46.47	52.39	59.40	**72.30**	38.21	40.75	44.63	**58.55**	32.47	34.47	37.32	**49.70**	29.33	30.69	33.72	**42.60**	28.21	29.18	31.37	**37.88**
Round 3	51.77	58.48	69.53	**83.00**	44.20	47.23	55.36	**70.40**	38.59	40.38	46.97	**61.03**	35.47	37.12	41.40	**53.55**	34.25	35.42	39.34	**48.24**

Table 1. Comparison of the four algorithms.

Baseline methods We compared our MV-BMAL with three baselines: SVM-AL [5], TSVM-AL [6] and Co-SVM [2]. SVM-AL and TSVM-AL take all features into one vector (single view) and use simply top-k active strategy. Co-SVM [2] considers the features from the color and texture views, but ignores the shape view, and it also adopts the top-k active strategy. Different from the baseline methods, our MV-BMAL based on SF-cuts (MV-SF-cuts) considers features from color, texture and shape views and adopts our proposed SF-cuts active strategy. In order to prove the effectiveness of our proposed SF-cuts, we compared it with MV-BMAL based on Top-k (MV-Top-k).

Performance Comparison. The experimental comparison results are illustrated in Fig. 2 and Tab. 1. We can observe that different training models (single view vs. multi-view) and different active learning strategies (top-k batch model and SF-cuts batch model) have a significant impact on retrieval performance. Single view ignores the differences of the contribution of each sub-features and top-k active strategy chooses feedback samples closest to the support vector boundary in which some samples could be similar or even identical to each other. This is the reason why our MV-BMAL method can greatly improve the performance compared to the baselines.

4. Conclusion

In this paper, we proposed a novel semi-supervised active learning framework (MV-BMAL) for improving the performance of image retrieval. The MV-BMAL divides the features into independent views considering the differences of the contribution of each sub-features. Meanwhile it adopts the proposed SF-cuts active strategy to select informative and representative feedback samples. Finally, the experimental results show that our MV-BMAL method makes a considerable improvement compared with the baseline algorithms.

5. Acknowledgement

The authors acknowledge the support by Natural Science Foundation of China (NSFC) No.60833004.

References

[1] J. C. Bezdek, R. Ehrlich, and W. Full. Fcm: The fuzzy c-means clustering algorithm. *Computers and Geosciences*, 10:191–203, 1984.

[2] J. Cheng and K. Wang. Active learning for image retrieval with co-svm. *Pattern Recognition*, 40:330–334, 2006.

[3] C. M. Christoudias, R. Urtasun, and T. Darrell. Multi-view learning in the presence of view disagreement. In *Proc. of UAI'08*, 2008.

[4] R. Datta, D. Joshi, and et. al. Image retrieval: ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, 2008.

[5] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proc. 9th ACM Multimedia Conference*, pages 107–118, 2001.

[6] L. Wang, K. L. Chan, and Z. Zhang. Bootstrapping svm active learning by incorporating unlabelled images for image retrieval. In *Proc. of CVPR'03*, volume 1, pages 629–634, 2003.

[7] R. Yong, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Trans. on CSVT*, 8(5):644–655, 1998.