Multi-modal Multi-label Semantic Indexing of Images Based on Hybrid Ensemble Learning

Wei Li[1], Maosong Sun[1], and Christopher Habel[2]

[1] State Key Lab of Intelligent Technology and Systems
Department of Computer Science and Technology, Tsinghua University
Beijing 100084, P.R. China
wei.lee04@gmail.com, sms@mail.tsinghua.edu.cn
[2] Fachbereich Informatik, Universität Hamburg
Hamburg, 22527, Germany
habel@informatik.uni-hamburg.de

Abstract. Automatic image annotation (AIA) refers to the association of words to whole images which is considered as a promising and effective approach to bridge the semantic gap between low-level visual features and high-level semantic concepts. In this paper, we formulate the task of image annotation as a multi-label multi class semantic image classification problem and propose a simple yet effective method: hybrid ensemble learning framework in which multi-label classifier based on uni-modal features and ensemble classifier based on bi-modal features are integrated into a joint classification model to perform multi-modal multi-label semantic image annotation. We conducted experiments on two commonly-used keyframe and image collections: MediaMill and Scene dataset including about 40,000 examples. The empirical studies demonstrated that the proposed hybrid ensemble learning method can enhance a given weak multi-label classifier to some extent, showing the effectiveness of our proposed method when limited number of multi-labeled training data is available.

1 Introduction

Automatic image annotation (AIA) refers to the association of semantic concepts to whole images which has become a hot research topic and increasingly required by many modern applications. For example, in the domain of semantic scene classification and medical image interpretation, multi-modal indexing through AIA enables each image or video clips to be associated with one or more descriptive concepts which allows for semantic browsing and retrieval of visual information via different keywords at different semantic levels when an ontology or concept hierarchy is available. Through the sustained efforts of experts and researchers, many approaches based on computer vision and machine learning theory have been proposed to attack this problem, which, in general, can be categorized into three major classes: generative models [1-4, 6-8, 13-18, 31]; discriminative approaches [5, 9-12,21, 27, 29-30] and search and mining-based annotation [25]. Some of these approaches have achieved the state-of-the-art performance and proved that automatic image annotation is an

H.H.S. Ip et al. (Eds.): PCM 2007, LNCS 4810, pp. 744–754, 2007.

effective solution to bridge the notorious semantic gap between low-level perceptual features and high-level semantic concepts. However, the key characteristic of automatic image annotation is that each image is usually assigned to multiple different semantic labels simultaneously instead of single label, because each image may contain multiple objects with different semantics. So multi-label classification model is more suitable than traditional single-label classifiers in that correlations between semantic labels can be incorporated rather than treating them as independent labels. In this case, label ambiguity or incompatibility can be avoided. For example, "*sky*" and "*ocean*" are more likely to co-occur than "*sky*" and "*computer*". Furthermore, multi-labeled training data is hard to obtain or create in large quantities which require large amount of human labeling effort, limited number of labeled training images can hardly represent the distribution of visual features for a concept of interest. Consequently, how to build accurate classification model using the limited multi-labeled image data to improve the annotation accuracy is becoming an important research issue.

The main contribution of this paper is three-fold: First, we formulate the task of image annotation as a multi-label, multi class semantic image classification problem under a joint classification framework called hybrid ensemble learning. Second, we review the multi-label learning approaches, and evaluate some of them on the image annotation task. Third, to enhance the annotation accuracy, single-label ensemble classifier based on bi-modal features is again fused to refine the classification results given by the multi-label classifier using the uni-modal features. To model the possible dependency between labels, correlations among labels obtained by using latent semantic indexing are incorporated into the bi-modal feature space. To the best of our knowledge, hybrid ensemble learning methods which integrate multi-label classifier based on uni-modal features and ensemble classifier based on bi-modal features into a joint classification model has not been carefully investigated in the domain of automatic image annotation.

This paper is organized as follows: Section 2 discusses related work. Section 3 first reviews the literature of multi-label learning and classification, and then describes the hybrid ensemble learning framework. Section 4 shows our experimental results and some theoretical analysis. Conclusions and future work are discussed in Section 5.

2 Related Work

Recently, many models using machine learning techniques have been proposed for automatic image annotation and retrieval. In general, these methods can be categorized into three classes: generative models, discriminative approaches as well as search and mining-based techniques.

Generative models:

$$P(l, v) = \sum_{s} P(l|s)P(v|s)P(s) \quad l \subseteq L, v \in V \tag{1}$$

where v dentoes the image data, l the subset of semantic concepts, s is the latent variable, L and V are concept lexicon and visual feature space respectively. By computing

the joint distribution of visual features and associated concepts, the hidden correlation between this two modalities can be found and then is applied to annotate new images. Representative works are [1-4][6-8][13-18][31], especially R. Zhang et al[18] has achieved the state-of-the-art performance, G. Carneiro et al[31] proposed to use M-ary labeling and ignore the hidden variable which can reduce the model complexity.

Discriminative approaches:

$$P(w|v) \qquad w \in L, v \in V \tag{2}$$

where w is a concept from L. Instead of joint modeling of semantic concepts and visual features, discriminative approaches treat each concept as a single class label and directly model the posterior probability of w given v. Some attractive works are [5][9-12][21][27]. Among them, K.Goh et al[10] and Cees G.M. Snoek[27] can provide better results. M. Bouttell et al[21] proposed the cross-training method to conduct multi-label scene classification and introduced some specific evaluation metrics.

In short, generative models can handle a large number of classes and class imbalance problem in some degree, but the model complexity is a major hurdle. While discriminative approaches are computationally efficient, however, they are unable to scale well to a large number of classes since it requires one model to be built for each class.

Search and Mining-based annotation:

Apart from annotation by learning, Wang et al. [25] proposed annotation by search and mining techniques which can not only makes use of web-scale images but also allows for unlimited vocabulary.

More recently, learning with unlabeled images has become an active research area due to fact that large amount of labeled training images is hard to obtain or create in large quantities while limited number of training images can hardly represent the visual distribution of target concepts and more information is contained in the large pool of unlabeled ones. Feng et al [30] and Song et al [29] introduced the use of co-training and combination of active learning together with semi-supervised ensembling to perform semantic annotation of images and video clips.

3 The Framework of Image Annotation Model

3.1 Formulation of Automatic Image Annotation

Given a training set of annotated images, where each image is associated with a number of semantic labels. We make an assumption that each image can be considered as a multi-modal document containing both the visual component and semantic component. Visual component provides the image representation in visual feature space using low-level perceptual features including color and texture, etc. While, semantic component captures the image semantics in semantic feature space based on textual annotations derived from a generic vocabulary, such as "*sky*", "*ocean*", etc. Automatic image annotation is the task of discovering the association model between visual and semantic component from such a labeled image database and then applying the association model to generate annotations for unlabeled images. More formally, let ID denote the training set of annotated images:

- $ID = \{I_1, I_2, \ldots, I_N\}$
- each image I_j in ID can be represented by the combination of visual features and semantic labels in a multi-modal feature space, i.e., $I_j = \{L_j; V_j\}$
- semantic component L_j is a bag of words described by a binary vector $L_j = \{l_{j,1}, l_{j,2}, \cdots, l_{j,m}\}$, where m is the size of generic vocabulary, $l_{j,i}$ is a binary variable indicating whether or not the i-th label l_i appears in I_j
- visual component V_j may be more complex due to a large variety of methods for visual representation, in general, it can also have the vector form $V_j = \{v_{j,1}, v_{j,2}, \ldots, v_{j,n}\}$, for patch-based image representation, i.e., image I_j is composed of a number of image segments or fixed-size blocks, each of them is described by a feature vector $v_{j,i}$, and n is the number of image components; for global image representation, $v_{j,i}$ only denotes a feature component and n is the dimension of selected feature space

For a given unseen image represented by v_u, the goal of automatic image annotation is to estimate:

$$l^* = \arg\max p(l|v_u), \quad l \subseteq L, v_u \in V \tag{3}$$

3.2 Underlying Theory of Multi-label Learning and Classification

In traditional classification problems, to reduce the model complexity, class labels are assumed to be mutually exclusive or independent from each other and each instance to be classified belongs to only one class. However, in the context of image annotation, it is natural that one image belongs to multiple classes simultaneously due to the richness of image content, causing the actual classes to overlap in the feature space. Furthermore, in most cases, it is quite hard and insufficient to describe the image content using only a keyword because image semantics is represented by both basic semantic entities in that image and the relationships between them. Consequently, multi-label learning is a more suitable and intuitive solution for automatic image annotation.

Multi-label learning refers to the problem where each example is associated with multiple different class labels simultaneously. It is now ubiquitous in real-world applications, e.g., text categorization [19][22][23], protein function prediction[26]. And in scene classification [21], if we treat every concept as a class label, each scene image may belong to several semantic classes, such as "*sky*" and "*clouds*". In all these cases, instances for training are each associated with a set of labels, and the task is to predict a candidate label set for the unseen instance.

An intuitive approach to solving multi-label problem is to decompose it into multiple independent binary classification problems (one per class). However, this kind of method suffers from many disadvantages. One is that it does not scale well to a large number of classes since a binary classifier has to be built for each class. Second, it does not consider the correlations between the different labels. Third, it may encoun-

ter imbalanced data problem when the minority classes are given only a few labeled training examples. Another group of approaches toward multi-label learning is label ranking which stems from preference learning. Instead of learning binary classifiers for each class, these approaches learn a ranking function from the labeled examples that order class labels for a given test example according to their relevance to the example. Compared to the binary classification approaches, the label ranking approaches are advantageous in dealing with large numbers of classes because only a single ranking function is learned.

3.3 Hybrid Ensemble Learning for Multi-modal Image Annotation

In this paper, we propose a two-stage joint classification framework called hybrid ensemble learning. The main idea is to train two classifiers, multi-class multi-label classifier at first stage and binary-class single-label ensemble classifier at second-stage using uni-modal and bi-modal features respectively. For a new, unseen image, the multi-label classifier is first used to predict the possible labels, and then the ensemble classifier is responsible for determining whether or not each predicted label is appropriate to describe the image semantics. To be more formal, let X be the image data, Y the finite set of predefined semantic labels and the size of Y is denoted by k. For multi-labeled classifier training, each training pair has the uni-modal form of (x, y), where $x \in \mathrm{X}$, $y \subseteq \mathrm{Y}$ While, for ensemble classifier, the training data is derived using a natural reduction of multi-labeled data to binary data. To be more specific, each example is mapped to a k binary-labeled bi-modal meta-examples of the form $((x, l, r), y[l])$ for all $l \in \mathrm{Y}$, where $y[l] = 1$ if $l \in y$ and -1 otherwise, r denotes the correlation between the label l and all the other labels. In this paper, the correlation among different labels is obtained by using latent semantic indexing. That is, the observation of each derived meta-example is (x, l, r) , and the associated binary label is $y[l] \in \{-1, 1\}$. For the classification of a new image, the multi-label classifier is initially applied and a label list containing candidate labels is output. Each candidate label is then appended to the feature vector of the new image to form the above-mentioned bi-modal meta-example; this meta-example is finally classified by the ensemble classifier to examine if each predicted label is relevant to the new, unseen image. In other words, the main task of the ensemble classifier is to conduct meta-example identification, to identify the positive and negative ones, then the appended label in the positive meta-example is considered as the correct label for the corresponding image and is kept in the predicted label list while the appended label in the negative one is removed from the predicted label list. Since in most multi-labeled image collections, the number of semantic labels for each image is rather small compared to the total number of predefined semantic labels, the produced bi-modal training data is extremely imbalanced in the sense that the number of negative meta-examples is much larger than that of positive meta-examples. To avoid the performance degradation of ensemble classifier due to the class imbalance problem, we propose to use the asymmetric bagging [24] to generate a classifier ensemble. The key idea behind asymmetric bagging is that keeping positive meta-examples the same for each base classifier and bootstrapping is only performed on the negative meta-examples to sample the same number as the positive meta-examples to construct a

balanced training set. To build a desired ensemble classification model, maximizing the diversity of each base classifier while maintaining the consistency with the training data is known to be an important goal, so in our method, each sampled negative subset is different from each other to ensure the diversity of training data. Moreover, logistic regression is used as the base classifier which requires less training time and low storage for built models compared to support vector machines [20]. In addition, logistic regression classifier has achieved the state-of-the-art performance in image classification tasks [28]. To further guarantee the performance of ensemble classifier, we use boosting method, AdaBoost, to enhance each base classifier. The following figure 1 and figure 2 show the joint classification framework and the asymmetric bagging algorithm.

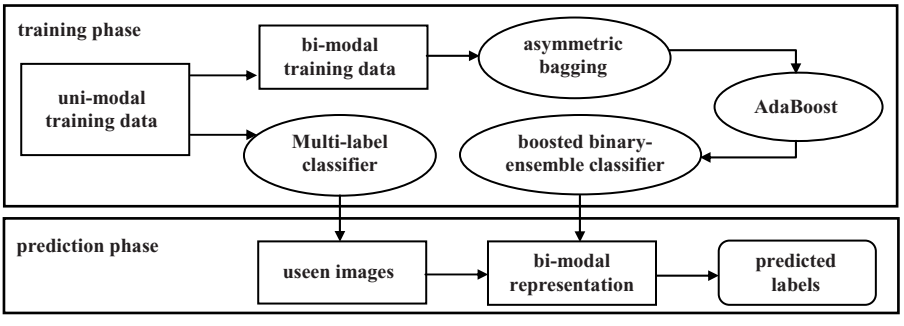

Fig. 1. Framework of Hybrid Ensemble Learning for Multi-Modal Image Annotation

Asymmetric Bagging Algorithm:

Input: positive meta-examples S^+, negative meta-examples S^-, base classifier I, number of base classifiers N, sampling factor α and the test meta-example t.

Output: final label l and classifier ensemble C

1. for i = 1 to N
2. S_i^- bootstrap samples from S^- using the criterion that $\alpha|S_i^-| = |S^+|$.
3. $I_i = I(S^+, S_i^-)$
4. $l = majority\_voting(I_i(x, S^+, S_i^-))$, $C = \{I_i\}$

Fig. 2. Algorithm of Asymmetric Bagging for Binary Classifier Training

4 Experimental Results

Data Set

Our experiments are carried out using two commonly-used keyframe and image datasets, MediaMill [27] and Scene [21] collection including about 42177 keyframes, 2407 images respectively. Table 1 shows the general information about the two data

collection. For the multi-label classifier, we use multi-label boosting [19] and multi-label C45[22] which have been successfully applied to text categorization tasks.

Mediamill: A number of color invariant texture features per pixel is firstly extracted. Based on these features, a set of predefined regions in a key frame image is labeled with similarity scores for a total of 15 low-level visual concepts. We vary the size of pre-defined regions to obtain a total of 8 concept occurrence histograms that characterize both global and local color-texture information. Finally, the histograms are concatenated to yield a 120-dimensional visual feature vector per keyframe image.

Scene: each image is divided into 49 blocks with the grid size of 7*7, then mean and variance of each block is computed in LUV color space, plus some computational inexpensive texture features, the resulting visual representation is 49 * 2 * 3 = 294 feature vector.

We here use the concepts of label cardinality and label density to describe the information of labels for each image. Let D be a multi-labeled image dataset including $|D|$ image pairs (x_i, y_i) and L the finite set of predefined semantic labels.

$$\textbf{label\_cardinality: } \frac{1}{|D|}\sum_{i=1}^{|D|}|y_i| \qquad \textbf{label\_density : } \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|y_i|}{|L|} \tag{4}$$

where label cardinality measures the average number of labels for each image and label density is the normalized representation of label cardinality.

Table 1. General Information of Two Datasets

Data Set	Examples		Feature Dimension	Labels	Label density	Label cardinality
	Training	Test				
MediaMill	29804	12373	120	101	0.0449	4.5369
Scene	1211	1196	294	6	0.1770	1.0619

Performance Metric

Multi-label Evaluation:

In contrast to traditional single-label classification, multi-label classification requires different evaluation metrics, here, we use the same metrics introduced in the literature. Let a multi-labeled image dataset denoted by D , which consists of $|D|$ image pairs (x_i, y_i) , L the lexicon of predefined semantic labels, y_i and z_i are the ground-truth and predicted label sequence respectively. In the following discussion, MLB and MLC45 refer to the multi-label boosting and multi-label C45 classifier. HMLB and HMLC45 with the suffix "H" refers to the boosted hybrid version of our method.

$$\textbf{Accuracy:}\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \qquad \textbf{Precision:}\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap Z_i|}{|Z_i|} \qquad \textbf{Recall:}\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap Z_i|}{|Y_i|} \tag{5}$$

Table 2. Multi-label Evaluation of Two Datasets

DataSet	Mediamill				Scene			
Methods	MLB	HMLB	MLC45	HMLC45	MLB	HMLB	MLC45	HMLC45
Accuracy	0.3897	0.3902	0.3020	0.3092	0.5074	0.5103	0.5152	0.5171
Precision	0.4621	0.4695	0.3893	0.3917	0.5114	0.5156	0.5371	0.5401
Recall	0.7262	0.7379	0.6117	0.6316	0.9511	0.9543	0.6442	0.6463

Retrieval Evaluation:

We also use the precision to evaluate the performance of the proposed method, for a single query concept *w*, precision is defined as follows. Let I_j denotes the retrieved *j*-th image, t_j and a_j represent the ground-truth labels and predicted labels associated with the *j*-th image.

$$\textbf{precision}(w) = \frac{\left|\left\{I_j \middle| w \in t_j \wedge w \in a_j\right\}\right|}{\left|\left\{I_j \middle| w \in a_j\right\}\right|} \tag{6}$$

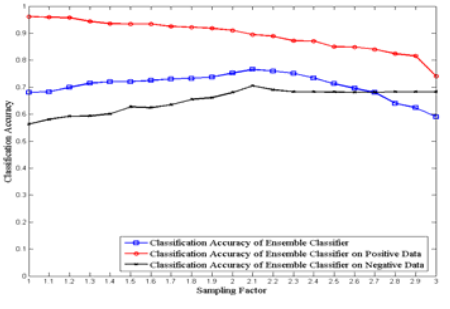

Fig. 3. Classification Accuracy vs Sampling Factor

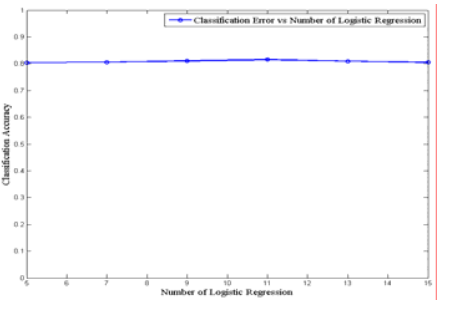

Fig. 4. Classification Accuracy vs Number of Logistic Regression

Figure 3 shows the classification accuracy of the ensemble classifier using different sampling factors in asymmetric bagging. We can see that using different sampling factor may lead to different classification accuracy. With the increasing of the sampling factor, classification accuracy of positive meta-examples may decreases while the classification accuracy of negative ones may increase, so we can find the best trade-off point to maximize the identification performance. In other words, maximizing the ability of discerning positive and negative meta-examples which ensures correctly predicted labels are kept in the candidate label list while the incorrectly predicted ones are removed.

Figure 4 illustrates the classification accuracy vs. number of logistic regression classifiers which verifies the fact that number of logistic regression has little effect on the classification accuracy of the ensemble classifier.

Table 3. Comparison of Precision using Different Methods

	concept	Training%	Test%	MLB	HMLB	MLC45	HMLC45		concept	Training%	Test%	MLB	HMLB	MLC45	HMLC45
1	aircraft	1.0267	0.986	0.1728	0.1868	0.0617	0.0617	56	maps	1.2012	1.2608	0.3333	0.3894	0.1131	0.125
2	allawi	0.218	0.0242	0	0	0	0	57	meeting	4.7141	5.0675	0.2446	0.2715	0.0898	0.073
3	anchor	5.2946	5.0594	0.3818	0.3962	0.3032	0.3197	58	military	4.3048	6.8698	0.2494	0.2501	0.1453	0.1567
4	animal	1.0368	0.9456	0.1304	0.1406	0.126	0.101	59	monologue	3.2278	2.4327	0.0902	0.129	0.0502	0.063
5	arrafat	0.6475	0.9133	0	0	0.0376	0.0264	60	motorbike	0.0537	0.1697	0	0	0	0
6	baseball	0.0134	0.4284	0	0	0	0	61	mountain	1.7045	1.0588	0.1029	0.134	0.0547	0.0612
7	basketball	0.7147	0.3556	0.1228	0.1344	0.0272	0.0317	62	natural_disaster	0.8388	0.9699	0.0781	0.0645	0.0501	0.0501
8	beach	0.0805	0.0647	0	0	0.0476	0	63	newspaper	0.3254	0.2829	0.3947	0.4148	0.1118	0.1208
9	bicycle	0.2113	0.0404	0	0	0.0149	0.0149	64	nightfire	0.1476	0.0566	0	0	0.0426	0.0589
10	bird	0.1879	0.2425	0.4118	0.4118	0.4286	0.4434	65	office	1.6273	1.8266	0.2727	0.2817	0.0418	0.0202
11	boat	0.812	0.5658	0.1014	0.1205	0.0512	0.0606	66	outdoor	33.989	40.006	0.5104	0.5363	0.5106	0.5364
12	building	7.1333	11.646	0.3198	0.3345	0.1924	0.2278	67	overlayed_text	37.784	35.796	0.4416	0.462	0.4463	0.4633
13	bus	0.4429	0.6708	0	0	0.0185	0.0185	68	people	80.764	79.189	0.8061	0.83	0.8375	0.8753
14	bush_jr	1.6743	0.5658	0.0811	0.0642	0.0136	0.0136	69	people_marching	2.0031	4.3078	0.3466	0.3324	0.1112	0.1268
15	bush_sr	0.208	0.0081	0	0	0	0	70	police_security	0.9596	0.8082	0	0	0.0223	0.0351
16	candle	0.0872	0.1051	0.1	0.1138	0	0	71	powell	0.047	0.493	0	0	0	0
17	car	5.0631	6.1909	0.2302	0.2529	0.1481	0.1667	72	prisoner	0.3456	0.2263	0	0	0.0056	0
18	cartoon	0.0872	0.2182	0.6	0.6328	0.2941	0.3024	73	racing	0.0906	0.1293	0	0	0	0
19	chair	0.6207	0.6062	0.2581	0.2789	0.1091	0.1258	74	religious_leader	0.2819	0.2344	0	0	0	0
20	charts	0.7851	0.5334	0.1597	0.171	0.0521	0.0332	75	river	0.1007	0.0889	0.1333	0.1501	0.0909	0.108
21	clinton	0.0503	0.2182	0	0	0.1333	0.1533	76	road	8.066	6.886	0.1974	0.2123	0.1326	0.1369
22	cloud	0.9059	1.6083	0.2576	0.2721	0.0709	0.0801	77	screen	1.5937	1.9801	0.08	0.061	0.049	0.0669
23	corporate_leader	2.6741	1.3578	0	0	0.0213	0.0213	78	sharon	0.0436	0.2021	0	0	0	0
24	court	0.2114	0.3152	0	0	0.0127	0.0127	79	sky	11.203	11.873	0.3547	0.3725	0.2564	0.2614
25	crowd	11.941	16.827	0.3856	0.3972	0.2726	0.3103	80	smoke	1.171	2.2387	0.3457	0.3526	0.1557	0.1709
26	cycling	0.1913	0.0323	0	0	0.0233	0.315	81	snow	0.4228	0.5496	0.0882	0.1081	0.0877	0.0935
27	desert	0.8388	1.5033	0.1887	0.1742	0.0402	0.0585	82	soccer	1.7347	0.3071	0.0577	0.0414	0.0641	0.0521
28	dog	0.1476	0.396	0.3	0.3218	0.0732	0.0607	83	splitscreen	0.8992	0.6223	0.25	0.28	0.0905	0.1072
29	drawing	0.0872	0.1778	0.5	0.4304	0	0	84	sports	3.9122	2.7237	0.105	0.1146	0.0716	0.0821
30	drawing_cartoon	0.1745	0.396	0.4167	0.4398	0.2083	0.2399	85	studio	14.206	14.823	0.4538	0.4628	0.4407	0.4525
31	duo_anchor	0.2751	0.1859	0.3077	0.3271	0.0441	0.0531	86	swimmingpool	0.0839	0.1051	0	0	0.0588	0.068
32	entertainment	20.427	13.101	0.1778	0.1935	0.1835	0.217	87	table	0.7751	0.5415	0.0543	0.0499	0.0341	0
33	explosion	0.5503	1.083	0.0962	0.0761	0.0562	0.0667	88	tank	0.0872	0.0808	0	0	0.0147	0
34	face	66.713	65.101	0.6983	0.7117	0.7365	0.7543	89	tennis	0.3523	0.5819	0.2143	0.2208	0.046	0.0557
35	female	4.5598	2.1983	0.1064	0.113	0.0402	0.0501	90	tony_blair	0.0503	0.2667	0	0	0	0
36	fireweapon	0.3624	0.5415	0.1429	0.1604	0.0332	0.0202	91	tower	0.7751	0.6547	0.0526	0.0433	0.0348	0.0412
37	fish	0.2785	0.1293	0.0741	0.0741	0.1389	0.1577	92	tree	0.8086	0.881	0.1032	0.1177	0.0747	0.0417
38	flag	1.3085	1.1719	0.2444	0.2628	0.0509	0.0625	93	truck	1.2112	1.0668	0.0543	0.0411	0.0383	0.0474
39	flag_usa	0.9563	0.9779	0.186	0.171	0.0539	0.0643	94	urban	12.25	9.1813	0.1969	0.1969	0.1374	0.1584
40	food	0.5234	0.8648	0.06	0.06	0.1908	0.2012	95	vegetation	4.0196	4.8412	0.1787	0.1911	0.1052	0.1241
41	football	0.2047	0.4041	0.1111	0.1264	0.0727	0.09	96	vehicle	7.9184	8.8984	0.2617	0.2751	0.1768	0.1974
42	golf	0.2617	0.3233	0	0	0.0577	0.0463	97	violence	8.3881	10.175	0.2899	0.3	0.197	0.2141
43	government_building	0.2852	0.194	0	0	0.0089	0	98	walking_running	14.156	17.571	0.2999	0.3114	0.2325	0.2431
44	government_leader	9.7269	8.2114	0.1956	0.2019	0.1132	0.1647	99	waterbody	2.4024	1.972	0.1344	0.1461	0.1335	0.1504
45	graphics	3.0097	3.6289	0.3162	0.3288	0.1631	0.1912	100	waterfall	0.0705	0.0808	0	0	0.1538	0.1752
46	grass	0.9361	0.6142	0.0494	0.0494	0.0379	0.041	101	weather	1.0301	1.3012	0.2917	0.3065	0.0878	0.0989
47	hassan_nasrallah	0.047	0.194	0	0	0	0								
48	horse	0.1711	0.0242	0	0	0	0								
49	horse_racing	0.1208	0.0242	0	0	0	0								
50	house	0.302	0.3799	0	0	0.0114	0.0157	1	beach	18.745	16.722	0.3848	0.3741	0.4605	0.4713
51	hu_jintao	0.0268	1.0749	0	0	0	0	2	fall foliage	13.625	16.639	0.6393	0.6508	0.7129	0.7483
52	indoor	20.376	22.129	0.4243	0.4485	0.415	0.4443	3	field	16.268	16.722	0.5246	0.5329	0.5481	0.5481
53	kerry	0.3053	0.0081	0	0	0	0	4	mountain	16.185	19.816	0.5158	0.5264	0.4821	0.5034
54	lahoud	0.312	0.1536	0.1429	0.1587	0.0676	0.0866	5	sunset	22.874	21.405	0.3377	0.3487	0.3558	0.3841
55	male	5.9388	2.4812	0.0924	0.1037	0.0607	0.0737	6	urban	18.497	17.308	0.3642	0.3714	0.4176	0.4299

Table 3 shows the precision results using different methods on Mediamill and Scene collection, which illustrates that our method can effectively remove the incorrectly predicted labels while keeping the correctly predicted ones. However, the performance of this method is dependent on the accuracy of the first-stage multi-label classifier; that is to say, we can not boost the zero-precision label outputted by the multi-label classifier. In addition, the precision value of concepts with sufficient training data is satisfactory in most cases, for example, "*people*", "*face*", etc. but, for the concept "*entertainment*" , its precision value is low, possible reasons are the large

variation of visual feature distribution of this concept. So it is hard to learn the visual patterns for some concept of interest.

5 Conclusion and Future work

In this paper, we propose a general framework for automatic image annotation and retrieval based on hybrid ensemble learning in which multi-label classifier based on uni-modal features and single-label ensemble classifier based on bi-modal features are integrated into a unified joint classification framework. Empirical results indicate that the advantage of our proposed method is that it can enhance the accuracy of a given mutli-label classifiers in some cases when limited number of multi-labeled training data is available. While the disadvantage is that its accuracy is dependent on the performance of the multi-label classifier, for example, our method has no effect on the zero-precision label in the test set. In addition, we can also draw other conclusions: First, in some cases, applying a sampling ratio factor to asymmetric bagging can lead to improved performance when majority-minority ratio is large. Second, the number of logistic regression classifiers does not affect the model performance.

Acknowledgements

We would like to express our deepest gratitude to Marcel Worring and Jiebo Luo for making their datasets available. The research is supported by the National Natural Science Foundation of China under grant number 60573187 and 60321002 and 60520130299.

References

1. Barnard, K., Dyugulu, P., de Freitas, N., Forsyth, D., Blei, D., Jordan, M.I.: Matching words and pictures. Journal of Machine Learning Research 3, 1107–1135 (2003)
2. Barnard, K., Forsyth, D.A.: Learning the Semantics of Words and Pictures. In: Proceedings of International Conference on Computer Vision, pp. 408–415 (2001)
3. Duygulu, P., Barnard, K., de Freitas, N., Forsyth, D.: Ojbect recognition as machine translation: Learning a lexicon fro a fixed image vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 97–112. Springer, Heidelberg (2002)
4. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proc. of SIGIR 2003, pp. 119–126 (2003)
5. Chang, E., Goh, K., Sychay, G., Wu, G.: CBSA: Content-based soft annotation for multimodal image retrieval using bayes point machines. IEEE Transactions on CSVT 13(1), 26–38 (2003)
6. Li, J., Wang, J.A.: Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Transactions on PAMI 25(10), 175–188 (2003)
7. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In: Proc. of the 16th Annual Conference on Neural Information Processing Systems (2004)
8. Blei, D., Jordan, M.I.: Modeling annotated data. In: Proceedings of the 26th intl. SIGIR Conf., pp. 127–134 (2003)

9. Li, B., Goh, K.: Confidence-based dynamic ensemble for image annotation and semantics discovery. In: Proc. of ACM MM 2003, pp. 195–206 (2003)
10. Goh, K., Li, B., Chang, E.: Semantics and feature discovery via confidence-based ensemble. ACM Transactions on Multimedia Computing, Communications, and Applications 1(2), 168–189 (2005)
11. Goh, K., Chang, E., Li, B.: Using on-class and two-class SVMs for multiclass image annotation. IEEE Trans. on Knowledge and Data Engineering 17(10), 1333–1346 (2005)
12. Fan, J., Gao, Y., Luo, H.: Multi-level annotation of natural scenes using dominant image components and semantic concepts. In: Proc. of ACM MM, pp. 540–547 (2004)
13. Feng, S.L., Lavrenko, V., Manmatha, R.: Multiple Bernoulli Relevance Models for Image and Video Annotation. In: Proc. of CVPR 2004 (2004)
14. Jin, R., Chai, J.Y., Si, L.: Effective Automatic image annotation via a coherent language model and active learning. In: Proc. of ACM MM 2004 (2004)
15. Kang, F., Jin, R., Chai, J.Y.: Regularizing Translation Models for Better Automatic Image Annotation. In: Proc. of CIKM 2004 (2004)
16. Monay, F., Gatica-Perez, D.: On image auto-annotation with latent space models. In: Proc. of ACM MM 2003. Conf. on Multimedia (2003)
17. Monay, F., Gatica-Perez, D.: PLSA-based image auto-annotation: Constraining the latent space. In: Proc. ACM Int. Conf. on Multimedia, New York (October 2004)
18. Zhang, R., Zhang, Z., Li, M., WY, M., Zhang, HJ.: A probabilistic semantic model for image annotation and multi-modal image retrieval. Multimedia Systems 12(1), 27–33 (2006)
19. Schapire, R., Singer, Y.: Boostexter: A boosting-based system for text categorization. Machine Learning 39, 135–168 (2000)
20. Wang, X.-R., Lin, C.-J.: LIBLR: a library for large regularized logistic regression (2007), Software available at http://www.csie.ntu.edu.tw/~cjlin/liblr/
21. Boutell, M., Luo, J., Shen, X., Luo, J.: Learning multi-label scene classification. Pattern Recognition 37(9), 1757–1771 (2004)
22. de Comite, F., Gilleron, R., Tommasi, M.: Learning multi-label alternating decision trees from texts and data. In: Proc. of MLDM 2003, pp. 35–49 (2003)
23. Gao, S., Wu, W., Lee, C.-H., Chua, T.-S.: A MFoM learning approach to robust multiclass multi-label text categorization. In: Proc. of ICML 2004, p. 42 (2004)
24. Tao, D., Xiaoou, T., Li, X., Wu, X.: Asymmetric Bagging and Random Subspace for Support Vector Machines-based Relevance Feedback in Image Retrieval. IEEE trans on PRMI 28(7), 1088–1099 (2006)
25. Wang, X., Zhang, L., Jing, F., Ma, W.-Y.: AnnoSearch: Image Auto-Annotation by Search. Proc. of CVPR (2006)
26. Chen, K., Lu, B.L., Kwok, J.T.: Effcient Classification of Multi-label and Imbalanced Data using Min-Max Modular Classifiers. In: Proc. of IJCNN 2006, pp. 1770–1775 (2006)
27. Snoek, C.G.M., Worring, M., van Gemert, J.C., Geusebroek, J.-M., Smeulders, A.W.M.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: Proc. Of ACM MM 2006, pp. 421–430 (2006)
28. Hoi, S.C., Jin, R., Lyu, M.: Batch Mode Active Learning and Its Application to Medical Image Classification. In: Proc. of ICML 2006, pp. 417–424 (2006)
29. Song, Y., Qi, G.-J., Hua, X.-S., Dai, L.-R., Wang, R.-H.: Video Annotation by Active Learning and Semi-Supervised Ensembling. In: Proc. of ICME 2006, pp. 933–936 (2006)
30. Feng, H., Chua, T.-S.: A bootstrapping approach to annotating large image collection. In: MIR 2003, pp. 55–62 (2003)
31. Carneiro, G., Chan, A.B., Moreno, P.J., Vasconcelos, N.: Supervised Learning of Semantic Classes for Image Annotation and Retrieval. IEEE trans on PAMI 29(3), 394–410 (2007)