A GENERALIZED DISCRIMINATIVE MUITIPLE INSTANCE LEARNING FOR MULTIMEDIA SEMANTIC CONCEPT DETECTION

Sheng Gao and Qibin Sun

Institute for Infocomm Research (I2R), A-Star, Singapore, 119613
{gaosheng, qibin}@i2r.a-star.edu.sg

ABSTRACT

In the paper we present a generalized discriminative multiple instance learning algorithm (GD-MIL) for multimedia semantic concept detection. It combines the capability of the MIL for automatically weighting the instances in the bag according to their relevance to the positive and negative classes, the expressive power of generative models, and the advantage of discriminative training. We evaluate the GD-MIL on the development set of TRECVID 2005 for high-level feature extraction task. The significant improvement is observed using the GD-MIL over the benchmark. The mean of AP's over 10 concepts using the GD-MIL is 4.18% on the validation set and 3.94 % on the evaluation set. As the comparison, they are 2.12% and 2.63% for the benchmark, correspondingly.

Index Terms – Multiple instance learning, multimedia semantic concept detection, discriminative training

1. INTRODUCTION

Enormous digital multimedia is archived and it is growing exponentially with the popularity of Internet and personal digital multimedia devices. However, efficiently managing (e.g. indexing, search and browsing) such giant multimedia database at the semantic level is still a challenge. In the past, extensive studies on content-based image retrieval have been worked on retrieving the image using the low-level feature similarity. Retrieving from the video stream at the semantic level attracts much attention in the recent years. The techniques are being advanced by the annual TREC video retrieval evaluation (*TRECVID*) organized by NIST[1]. They exploit the learning algorithms developed in machine learning and pattern recognition for detecting multimedia semantic concept. For each interested semantic concept, the training samples (i.e. key frames) are manually annotated based on the visual content. A keyframe is labeled as the positive class if it is relevant with the concept. Otherwise, it is the negative [3]. Then the supervised learning algorithms are applied to train a classifier based on the annotated training set. Finally, the classifier is used scoring and sorting the video shots.

In semantic concept detection, the label of the keyframe is a weak annotation for image content. It means that the semantic annotation is given at the image level, and is not at the regions or objects that are exactly relevant with the concept. For example, an image annotated as *Car* may have other objects such as *building*, *tree*, etc. Thus the positive training image has some regions irrelevant with the concept. They are the noises for training the positive concept model. Ideally, these parts should be removed from the image representation. However, annotating at the regions is time consuming. Most of the available data is labeled at the image level. It is interesting to develop a learning algorithm that can automatically de-emphasize the negative regions in the training stage.

Multiple instance learning (*MIL*) is such a framework [6]. It learns the classifier from the weaker annotation, and has been successfully applied to content-based image retrieval, image classification, and semantic concept detection [1, 4, 5, 7]. MIL estimates the positive target through the diverse density (*DD*) [6] or Expectation-Maximization diverse density (*EM-DD*) algorithm [8]. But the existing target estimation in MIL is based on the maximum likelihood (*ML*). It is not a discriminative training. As we know, discriminative training technique can improve robustness of the classifiers and play a crucial role on classification and information retrieval. Another drawback is that it does not incorporate the expressive power of generative models widely used in classification. In [4], a general MIL is proposed to tackle this issue. But they estimate the model parameters using the *ML* without discriminative training.

In the paper we present a generalized discriminative multiple instance learning algorithm (GD-MIL) to address the above two shortcomings in the conventional MIL. The GD-MIL fuses the capability of MIL for automatically weighting the regions according to their relevance to the positive and negative classes, the expressive power of generative models, and the advantage of discriminative training. We evaluate the proposed algorithm on the high-level feature extraction task based on the development set of TRECVID 2005.

[1] http://www-nlpir.nist.gov/projects/tv2005/tv2005.html

1-4244-0481-9/06/$20.00 ©2006 IEEE

In the next section, we first give a brief introduction about the conventional MIL. Then in Section 3, we discuss our GD-MIL approach for estimating the classifier for multimedia semantic concept detection. In Section 4, the experimental results and analysis are presented. Finally, we summarize our findings in Section 5.

2. MULTIPLE INSTANCE LEARNING

In MIL, the training samples are provided at the *bag* level, each bag containing multiple *instances* [6]. But the annotation is given for the *bag* and the instance labels are hidden. For our current task of multimedia semantic concept detection, the bag is the whole image while the instances are its regions. An instance is represented by a *D*-dimensional feature vector, and the union of instances in a bag describe the bag. A bag is annotated as the positive if one of its instances is relevant with the concept. Otherwise, the bag is annotated as the negative. We use the similar notations as in [6]. B_i^+ is the *i-th* positive bag whose size is $\left|B_i^+\right|$, and its *j-th* instance is B_{ij}^+. B_i^- is the *i-th* negative bag with the size $\left|B_i^-\right|$ and its *j-th* instance being B_{ij}^-. We assume there are *M* positive bags and *N* negative bags in the training set. The target, *t*, being estimated is a single or multiple points at the *D*-dimensional feature space.

The target, t^*, is estimated by maximizing the joint probability of the training samples defined in Eq. (1),

$$t^* = \max_t P\left(B_1^+,\cdots,B_M^+,B_1^-,\cdots,B_N^-\left|t\right.\right) \qquad (1).$$

Assuming the bags and instances are independently sampled and a uniform prior over the target, Eq. (1) will be,

$$t^* = \max_t \prod_i^M P\left(t\left|B_i^+\right.\right)\prod_i^N P\left(t\left|B_i^-\right.\right) \qquad (2).$$

It is the general definition of maximum diverse density algorithm [6]. To define the conditional probability in Eq. (2), a *noise-or* model is applied. The probability of a positive bag is defined as,

$$P\left(t\left|B_i^+\right.\right) = 1-\prod\nolimits_j\left(1-P\left(t\left|B_{ij}^+\right.\right)\right) \qquad (3).$$

Correspondingly, the probability of the negative bag is,

$$P\left(t\left|B_i^-\right.\right) = \prod\nolimits_j\left(1-P\left(t\left|B_{ij}^-\right.\right)\right) \qquad (4).$$

The probability of an instance in Eqs. (3-4) is calculated as,

$$P\left(t\left|B_{ij}^c\right.\right) = \exp\left(-\left\|B_{ij}^c - t\right\|^2\right), c=\left\{+,-\right\} \qquad (5).$$

With the definitions of Eqs.(3-5), the target can be found by maximizing Eq. (2) using the gradient descent algorithms [6, 8].

Eq. (5) is a single Gaussian distribution model of the instance with the target as its mean. Because the same model is used for the instance of positive bag and that of negative, the estimation by optimizing Eq. (2) lacks discrimination.

3. GENERALIZED DISCRIMINATIVE MULTIPLE INSTANCE LEARNING

The generalized discriminative multiple instance learning algorithm (GD-MIL) is introduced to provide a MIL framework so that 1) the generative model is easy to be incorporated, and 2) the model robustness is improved by discriminative training. The GD-MIL can be used for any generative model and is easy to extend to multi-class classification. But here we will use the Gaussian mixture model (*GMM*) as an example to discuss the GD-MIL and its estimation for the binary classification or detection problem.

3.1. Gaussian Mixture Model

For the positive or negative class, the GMM is used to model their feature distribution of the instance. The GMM's are defined as,

$$g^c\left(x\left|w^c,\mu^c,\Sigma^c\right.\right) = \sum\nolimits_{k=1}^{K^c} w_k^c \cdot N\left(x\left|\mu_k^c,\Sigma_k^c\right.\right), c=\left\{+,-\right\} \qquad (6).$$

In the above, *x* is the instance feature vector, and the superscript, *c*, notes the class (+: *positive*, -: *negative*). K^c is the mixture number, w_k^c is the weight coefficients, and *N(.)* is the Gaussian distribution with the mean μ_k^c and covariance matrix Σ_k^c (Here the diagonal matrix is used). Thus the parameter set is $\Lambda = \left\{w_k^c,\mu_k^c,\Sigma_k^c\right\},\ k\in\left[1,K^c\right]$.

3.2. Generalized Likelihood Ratio

As the models in Eq. (6) are known, we can determine whether an instance, x_i, in the bag $X=\left(x_1,\cdots,x_{|X|}\right)$ is a positive or negative according to the following decision rule,

$$c_i = \begin{cases} Positive, if\ \ l\left(x_i\left|\Lambda\right.\right) > th \\ Negative, \quad\ Otherwise \end{cases} \qquad (7),$$

with c_i being the assigned label for the instance x_i, and *th* being a threshold (here it is set zero), and

$$l\left(x_i\left|\Lambda\right.\right) = \log g^+\left(x_i\left|w^+,\mu^+,\Sigma^+\right.\right) - \log g^-\left(x_i\left|w^-,\mu^-,\Sigma^-\right.\right) \qquad (8).$$

Eq. (8) is the log-likelihood ratio (*LLR*) between the positive and negative classes. To apply the idea of MIL, we further define a function as in Eq. (9) so that a positive or negative label, *C(X)*, is assigned to the bag, *X*.

$$C\left(X\right) = \begin{cases} Positive,\ if\ \ \max_i l\left(x_i\left|\Lambda\right.\right) > th \\ Negative, \qquad\qquad Otherwise \end{cases} \qquad (9),$$

i.e. the bag label is decided according to the maximum LLR over all instances in the bag. If the maximum LLR is more than the threshold, the bag is the positive. Otherwise, it is the negative. It is alternate definition for the *noise-or* model, similar as in [9] when extending SVM to handle the MIL.

However, Eq. (9) is not smoothing and differential. To address the issue, a generalized likelihood ratio (*GLR*) is defined in Eq. (10) over a bag to approximate Eq. (9),

$$f\left(X\left|\Lambda\right.\right) = \frac{1}{\eta}\log\left(\frac{1}{\left|X\right|}\sum\nolimits_{j=1}^{|X|}\exp\left(\eta\cdot l\left(x_i\left|\Lambda\right.\right)\right)\right) \qquad (10).$$

η is a positive constant. Eq. (10) is equal to Eq. (9) as η is the positive infinite. A bag is assigned as a positive when Eq. (10) is more than the threshold. More interestingly, Eq. (10) is a differential function over the model parameters.

The GLR in Eq. (10) combines the evidence of each instance in the bag so that the final decision at the bag is determined. It bridges the instance-level evidence and the bag-level label.

3.3. GD-MIL Algorithm

As the GLR is calculated, the probability of a positive bag, B_i^+, is approximated using a sigmoid function,

$$P\left(c=+\left|B_i^+,\Lambda\right.\right)=\frac{1}{1+\exp\left(-\alpha\cdot f\left(B_i^+\left|\Lambda\right.\right)\right)} \quad \text{(11a).}$$

Thus the probability of a negative bag, B_i^-, is calculated as,

$$P\left(c=-\left|B_i^-,\Lambda\right.\right)=1-\frac{1}{1+\exp\left(-\alpha\cdot f\left(B_i^-\left|\Lambda\right.\right)\right)} \quad \text{(11b).}$$

With the above definitions, we can formulate the objective function being optimized as the log-likelihood function of Eq. (2) over the training set, i.e.,

$$L=\sum_{i=1}^{M}\log P\left(c=+\left|B_i^+,\Lambda\right.\right)+\sum_{i=1}^{N}\log P\left(c=-\left|B_i^-,\Lambda\right.\right) \quad \text{(12).}$$

Eq. (12) is the objective function maximized in our GD-MIL. Because the likelihood ratio between the positive and negative models is embedded into the object function in Eq. (12), the model estimation is discriminative training. It means that both of the positive and negative samples can contribute to the parameter estimation, unlike the maximum likelihood algorithm, where the class model is trained only from itself samples.

The gradients with respect to the parameters, $\Lambda=\left\{w_k^c,\mu_k^c,\Sigma_k^c\right\}$, are easily derived from Eq. (12), and are shown in the following for the *(t+1)-th* iteration.

$$\begin{aligned}\nabla L\Big|_{\Lambda_{t+1}}=\alpha\Big(&\sum_{i=1}^{M}\left(1-P\left(c=+\left|B_i^+,\Lambda_t\right.\right)\right)\cdot\Delta f\left(B_i^+\left|\Lambda\right.\right)\Big|_{\Lambda_t} \\ &-\sum_{i=1}^{N}\left(1-P\left(c=-\left|B_i^-,\Lambda_t\right.\right)\right)\cdot\Delta f\left(B_i^-\left|\Lambda\right.\right)\Big|_{\Lambda_t}\Big)\end{aligned} \quad \text{(13),}$$

with,

$$\nabla f\left(B_i^c\left|\Lambda\right.\right)\Big|_{\Lambda_t}=\sum_{j=1}^{\left|B_i^c\right|}b_j^c\cdot\left(\nabla\log g^+\left(B_{ij}^c\left|\Lambda\right.\right)\Big|_{\Lambda_t}-\nabla\log g^-\left(B_{ij}^c\left|\Lambda\right.\right)\Big|_{\Lambda_t}\right) \quad \text{(13a),}$$

and

$$b_j^c=\exp\left(\eta\cdot l\left(B_{ij}^c\right)\right)\Big/\sum_{k=1}^{\left|B_i^c\right|}\exp\left(\eta\cdot l\left(B_{ik}^c\right)\right) \quad \text{(13b).}$$

The b_j^c's in the above weight each instance in a bag. Eq. (13b) describes that the heavy weight is given to the instance with strong discrimination, i.e. the instance with higher LLR. We skipped the details of the gradients for GMM in Eq. (13a).

When the gradients in Eq. (13) are gotten, the iterative algorithm is used to find the optimal model parameters,

$$\Lambda_{t+1}=\Lambda_t+k\cdot\nabla L\Big|_{\Lambda_t} \quad \text{(14)}$$

k controls the learning rate.

3.4. Adjusting Control Parameters

The control parameters, η, α and *k*, are empirically set as, $\eta=5.0$, $\alpha=0.5$, $k=0.2$ for all concepts based on our preliminary results. We cannot find significant performance difference when adjusting η and α around the above setting. The learning rate should be set as: 1) it is bigger enough so that the log-likelihood in Eq.(12) monotonically convergences to its maximum.

4. EXPERIMENTAL RESULTS AND ANALYSIS

We analyze the presented GD-MIL algorithm on the development set of TRECVID 2005 for evaluating semantic concept detection task. The positive and negative GMM's are trained using the algorithm introduced in Section 3. The benchmark is trained without MIL, i.e., train positive GMM model on all instances in the positive bags and negative model on all instances in the negative bags. We also implement the GMIL presented in [4]. But we find it is comparable with our benchmark on our dataset and image representation method.

Table 1 Description of the TRECVID 2005 dataset

	T	V	E
Building	41,978 (3,604)	11,173 (1,416)	8,295 (1,064)
Car	41,919 (2,253)	11,325 (767)	8,487 (370)
Explosion	42,038 (641)	11,301 (81)	8,497 (26)
Flag-US	42052 (337)	10,970 (51)	8,497 (92)
Maps	41,988 (594)	11,290 (171)	8,473 (145)
Mountain	42,073 (385)	11,331 (168)	8,496 (73)
People	42,021 (996)	11,321 (221)	8,473 (91)
Prisoner	42,003 (61)	11,332 (43)	8,112 (2)
Sports	41,753 (1,140)	11,310 (295)	8,498 (135)
Waterscape	42,043 (819)	11,312 (152)	8,484 (110)

4.1. Evaluation Metrics

We compare the results using the non-interpolated average precision (AP), an official metric for TREC,

$$AP=\frac{1}{R}\sum_{i=1}^{Q}\frac{R_i}{i}*I_i\,. \quad \text{(15)}$$

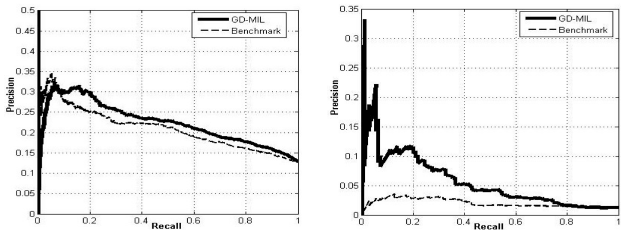
R is the number of true relevant image documents in the evaluation set. Q is the number of retrieved documents by the system (Here Q=2000 same as used in TRECVID official evaluation). I_i is the *i-th* indicator in the rank list with Q images. It is 1 if the *i-th* image is relevant and zero otherwise. R_i is the number of relevant in the top-i images.

4.2. Experimental Setup

The development set of TRECVID 2005 has 74,509 keyframes extracted from 137 news videos (~80 hours). 10 concepts are used for official evaluation. We divide the set into 3 parts for training, validation, and evaluation, respectively. The visual feature is the 12-dimensional texture (energy of log Gabor filter) extracted from a 32x32 grid. Other expressive visual features will be evaluated in the future. But here we concern the efficiency of the proposed learning algorithm rather than feature extraction. Each keyframe is uniformly segmented into 77 grids. The dataset is detailed in Table 1 (Column *T*: training set, Column *V*: validation set, Column *E*: evaluation set). The positive sample size is shown in the parentheses. For example, in the first row and column *T*, 41,978(3,604) means that there are 41,978 images in the training set for the concept *Building*, in which there are 3,604 images labeled as *Building*.

Table 2 AP values (%) and the number of relevant keyframes @ 2000 (*Pos.#*) for GD-MIL(Column *GD-MIL* and benchmark (Column *REF*) (Pos.#: the former number is for GD-MIL and the latter is for the benchmark)

Class	V			E		
	GD-MIL	REF	Pos.#	GD-MIL	REF	Pos.#
Building	**7.47**	5.63	449/399	**11.88**	10.89	469/444
Car	**3.93**	2.26	250/194	**6.34**	1.93	188/111
Explosion	**0.58**	0.25	26/21	**0.41**	0.40	13/15
US_Flag	**0.13**	0.04	9/7	**0.84**	0.83	37/34
Maps	**3.72**	0.68	67/37	**5.61**	4.97	114/104
Mountain	**6.09**	2.45	111/70	**3.93**	0.90	56/36
People	**2.05**	2.03	79/82	1.66	**4.02**	44/51
Prisoner	0.27	**0.97**	15/18	0.0	0.0	0/0
Sports	**5.47**	2.29	148/110	**3.67**	1.18	87/41
Waterscape	**12.12**	4.55	95/80	**5.08**	1.17	63/45
Avg.	**4.18**	2.12		**3.94**	2.63	



(a) Building (b) Waterscape

Figure 1 Precision-Recall curves for 2 concepts, *Building (a)* and *Waterscape* (b) (GD-MIL: *thick solid line*. Benchmark: *thin dash line*).

4.3. Experimental Results

The single Gaussian GMM models are used in all experiments. The experimental results on the validation and evaluation sets over 10 concepts are shown in Table 2. The best results for each concept on each set are highlighted. The GD-MIL (Column *GD-MIL*) performs better than the benchmark (Column *REF*) on 9 concepts out of 10. The average AP value over 10 concepts is 4.18% on the validation set and 3.94% on the evaluation set for the GD-MIL. As the comparison, the benchmark only has AP values 2.12% and 2.63%, correspondingly. Thus singnificant improvements in terms of AP are obtained.

After the MIL training, we expect the learned model is more *pure*. We mean the positive model has less affection from the negative instances in the positive bags. So we expect it will help improve the recall. The column *Pos.#* in Table 2 gives the number of relevant keyframes at the top-2000. Obviously, the GD-MIL retrieves more relevant docuemnts than the benchmark.

The AP only gives a coarse view of the classifier performance. For more details at every operating point, the Precision-Recall (*PR*) curve is used for evaluating the classifier. We plot the PRcurves in Figure 1 for two selected conceopts, i.e. *Building* and *Waterscape*, so that we can get a overall view of the GD-MIL. Once again, the improvements using the GD-MIL are obviously seen.

5. CONCLUSION

In the paper we present a generalized discriminative multiple instance learning algorithm (GD-MIL) for multimedia semantic concept detection. It combines the capability of the MIL for automatically weighting the instances in the bag according to their relevance to the positive and negative classes, the expressive power of generative models, and the advantage of discriminative training. We evaluate the GD-MIL on the development set of TRECVID 2005 for high-level feature extraction task. The significant improvement is observed using the GD-MIL. The mean of AP's over 10 concepts using the GD-MIL is 4.18% on the validation set and 3.94% on the evaluation set. As the comparison, they are 2.12% and 2.63% for the benchmark, correspondingly. In the future, we will evaluate the GD-MIL on more image features, and give a systematically comparison with other existing MIL algorithms.

6. REFERENCES

[1] C. Yang & T. Lozano-Perez, "Image database retrieval with multiple-instance learning techniques," *Proc. of ICDE'00.*

[2] J. Ramon & L. D. Raedt, "Multi instance neural network," Proc. of workshop at ICML'00 on Attribute-Value and Relational Learning: Crossing the boundaries, 2000.

[3] M. Naphade, et al., "A light scale concept ontology for multimedia understanding for TRECVID 2005," IBM Research Report RC23612 (W0505-104), May, 2005.

[4] M. Naphade & J.R. Smith, "A generalized multiple instance learning algorithm for large scale modeling of multimedia semantics," *Proc. of ICASSP'05.*

[5] O. Maron & A. L. Ratan, "Multiple-instance learning for natural scene classification," *ICML'*98.

[6] O. Maron & T. Lozano-Perez, "A framework for multiple instance learning," *Neural Information Processing Systems,* 1998.

[7] Q. Zhang, et al., "Content-based image retrieval using multiple –instance learning," *ICML'02.*

[8] Q. Zhang & S.A. Goldman, "EM-DD: an improved multiple-instance learning technique," *Neural Information Processing Systems,* 2001.

[9] S. Andrews, I. Tsochantaridis, & T. Hofmann, "Support vector machines for multiple-instance learning," *Neural Information Processing Systems,* 2004.