

HIDDEN MAXIMUM ENTROPY APPROACH FOR VISUAL CONCEPT MODELING

Sheng Gao, Joo-Hwee Lim and Qibin Sun

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore, 119613

{gaosheng, joohee, qibin}@i2r.a-star.edu.sg

ABSTRACT

Recently, the bag-of-words approach has been successfully applied to automatic image annotation, object recognition, etc. The method needs to first quantize an image using the visual terms and then extract the image-level statistics for classification. Although successful applications have been reported, it lacks the capability to model the spatial dependency and the correspondence between the patches and visual parts. Moreover, quantization deteriorates the descriptive power of patch feature. This paper proposes the hidden maximum entropy (HME) approach for modeling visual concepts. Each concept is composed of a set of visual parts, each part having a Gaussian distribution. The spatial dependency and image-level statistics of parts are modeled through the maximum entropy. The model is learned using the developed EM-IIS algorithm. We report the preliminary results on the 260 concepts in the Corel dataset and compared with the maximum entropy (ME) approach. Our experiments on concept detection show that (1) a relative increment of 10.3% is observed when comparing the average AUC value of HME approach with that of the ME approach and (2) the HME approach reduces the average equal error rate from 0.412 for the ME approach to 0.354.

1. INTRODUCTION

The bag-of-words approach is commonly adopted in text information retrieval so that a text document is represented in terms of word occurrences [2]. Although the method ignores syntactic and semantic information, it has been shown to be successful in text categorization and retrieval. Recently, it attracts increasing attention when dealing with image-related classification problems such as automatic image annotation, scene classification, object recognition, etc [1, 3-11]. Unlike text document, image document is not symbolic. The clustering algorithms (e.g. k-means) are first utilized to quantize the images into clusters, a discrete equivalent of the symbolic tokens in text documents. A high-dimensional feature vector is then extracted using the feature extraction techniques for text. Finally, machine learning algorithms, e.g., SVM [4], MC MFoM [10], CMRM [6], ME [11], are exploited to train the concept model. Good performances have been reported on the tasks such as image annotation and object recognition.

The bag-of-words approach makes it easy to utilize the image-level statistics. The quantities, such as unigram and bigram of visual terms, are invariant to rotation and robust to partial occlusion. This is one reason for its success in visual concept classification. However, the approach lacks the capability to model the spatial dependency and the correspondence between the patches and concepts. For some concepts, the spatial configuration may be informative for discriminating them from others. In addition, quantization loses some discriminative information held in continuous visual features. The loss is not recovered, regardless of how powerful the image model is.

To utilize the spatial configuration of patches in the visual concept model and to discover the correspondence between the patches and concept parts, the generative object model is an

attractive option [9]. Each concept is composed of a few visual parts, each visual part having a Gaussian distribution. The correspondence is treated as a hidden random variable and is learned using the EM algorithm. The generative model makes it feasible to integrate the appearance, shape and local spatial dependency. But it cannot capture the image-level statistics of parts because the computation cost increases exponentially with the order of spatial dependency.

In the paper, a hidden maximum entropy (HME) approach is presented for modeling visual concepts so that we can obtain the benefits from both the bag-of-words and the generative models and can address their each individual drawback. The HME models the concept using the visual parts, each part being a Gaussian distribution. The parts connect the patch feature with the discrete symbols, i.e., parts. The part configuration in the image and their interaction are further modeled through the ME model. Since the correspondence between the patches and the parts is unknown, an EM-IIS algorithm, i.e., EM embedded with improved iterative scaling (IIS), is developed.

The HME directly characterizes the distributions of the patch feature and part configuration. Thus, quantization is not necessary. When the correspondence is deterministic and all concept models share the visual parts, the HME would become the ME. It will be detailed in section 3.

Similar to the part-based object model (e.g. [9]), the HME approach utilizes the generative visual part models to characterize the patch feature distributions. But their difference is obvious. In [9], the hidden variable is the object part configuration, which describes the best matching between the parts and the patches. Thus, for a fixed configuration, each object part occurs at most once. While in the HME model, the hidden variable is the patch configuration of an image. It describes the best possible part that can generate a patch. For a fixed configuration, each part may occur many times. Moreover, the HME models the image-level statistics through the feature extractors and maximum entropy model which is infeasible using the model in [9]. We observed that the most relevant work is the latent maximum entropy (LME) model [12], which addresses learning with missing variables using the maximum entropy principle. But the learning algorithm is inefficient when the image-level statistics are used because the interactions among the parts are non-linear. In this paper, we present a feasible learning algorithm, i.e., EM-IIS (improved iterative scaling), to estimate the HME based concept models.

The paper is organized as the following. In the next section, we discuss the ME based concept model. Then the HME model is introduced in Section 3. The experimental results are reported in Section 4. Finally, we summarize our findings in Section 5.

2. MAXIMUM ENTROPY BASED CONCEPT MODELING

The maximum entropy model has been applied to text documents. However, the natural representation of image is a set of patch features (e.g., color, texture, etc.). Thus, the first step is to learn visual terms using the clustering algorithms such as k-means. Then a patch is quantized into its closest visual term based on its distance to the centroids of visual terms.

Quantization makes an image document to be a symbolic document. Any technique, including feature extraction and modeling algorithms, developed for text categorization can then be applied. Among them, the ME model is promising because of its capability to fuse diverse features.

Assuming there are M visual concepts and a set of training samples, $T = \{(I_i, y_i)\}$, where (I_i, y_i) is an image document I_i and its annotated concept, y_i , then M concept models are trained. The extracted patch features are denoted as, $I_i = (x_i(1), x_i(2), \dots, x_i(L_i))$, with $x_i(\cdot)$ being a D -dimensional vector and L_i being the number of vectors. After quantization using K visual terms, the image is represented as, $I_i = (q_i(1), q_i(2), \dots, q_i(L_i))$, where $q_i(\cdot)$ is a visual term quantized for $x_i(\cdot)$. From the quantized image document, the text-like features, e.g. *tf-idf*, *unigram*, *bigram*, etc., can be extracted [2]. For the ME model based classification, the popular feature extractors extract *unigram* or *bigram* features (e.g., [5, 11]).

2.1 Maximum Entropy Model

For M concept classification, we would like to calculate the predicted conditional probability, $P(y|I, \theta)$, for the concept y , where $y \in [1, M]$, I is the symbolized image document and θ is the model parameter set. Then the image will be assigned to the concept, y^* , which has the maximal predicted probability, i.e.,

$$y^* = \arg \max_{y \in [1, M]} P(y|I, \theta) \quad (1)$$

Assuming N feature extractors, $f_i(I, y)$, are designed to extract informative features, then the ME model can be estimated by maximizing the empirical maximum entropy in the training set under a set of constraints. These constraints state that the empirical quantity of each feature in the training set must be equal to its predicted value by the learned model. The optimization criterion gives rise to the following form of the visual concept model (Refer to [13] for details),

$$P(y|I, \theta) = \frac{1}{Z(I, \theta)} \exp\left(\sum_i \lambda_i \cdot f_i(I, y)\right) \quad (2)$$

where $Z(I, \theta) = \sum_y \exp\left(\sum_i \lambda_i \cdot f_i(I, y)\right)$, $\theta = \{\lambda_1, \lambda_2, \dots, \lambda_N\}$.

Similar to [4, 5, 11], the unigram feature is used in the paper. It is defined as,

$$f_{q,c}(I, y) = \begin{cases} \frac{\#(q, I)}{|I|}, & \text{if } c = y \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where q is a visual term, c is a visual concept and $|I|$ is the occurrence number of all terms in I . Now the feature extractor is indexed by both q and c . Eq. (3) implies each concept model in Eq. (2) has its own distinct parameters.

Maximizing the log-likelihood in the training set gives rise to the model parameters. Generalized iterative scaling (GIS) or IIS algorithm is developed for efficient estimation [13].

2.2 Limitations of the ME Model

With carefully designed feature extractors, the ME based concept model can capture the image-level statistics of visual terms, e.g. occurrence of a visual term or co-occurrence of any two visual terms. These features are invariant to rotation and are

robust to partial occlusion. However, the spatial dependency of patches is not characterized. It also cannot represent the correspondence between the image patches and the visual parts of the concepts because of its image-level representation. Moreover, quantization error is unavoidable in the method and it cannot be recovered by the subsequent process. Therefore, the bag-of-words description limits the utilization of patch features. For instance, currently only the appearance feature is used to generate the visual terms. It is challenging to incorporate the shape model because the shape model needs the spatial configuration of the concept parts.

In the next section, we will introduce the hidden maximum entropy approach to address the issue. The HME model treats learning visual terms and estimating the correspondence as a hidden stage. It jointly characterizes the distribution of patch features and part configuration of visual concept.

3. HIDDEN MAXIMUM ENTROPY CONCEPT MODELING

The bag-of-words based visual concept models consist of a universal visual term models, which map the image patches to their closest visual terms, and a concept dependent model, which characterizes the image-level statistics of the symbolized image. While in the HME based visual concept model, each concept model may have their distinct visual parts (hereafter, we don't distinguish visual part from visual term) which generate the observed patch features, and the image-level statistics for one correspondence is described by the ME model. The HME is different from the bag-of-words approach in that the correspondence mapping in HME is hidden and probabilistic while in the bag-of-words approach, it is deterministic.

3.1 Hidden Maximum Entropy Principle

For the image patch representation, $I = (x(1), x(2), \dots, x(L))$, and the concept models, there is a hidden random variable, $H = (h(1), h(2), \dots, h(L))$, to describe the mapping between the concept visual parts and the patches. Here $h(i)$ is one of K parts. Thus the log-likelihood for predicting the concept y is calculated through summing all possible mappings, i.e.,

$$\log(P(y|I, \theta)) = \log \sum_H P(y, H|I, \theta) \quad (4)$$

For the K -parts concept model and L -patches image, there will be K^L correspondence. Thus, the computation cost for the above sum is very huge. Even it is possible, there is another challenge to find a computationally tractable function for the joint distribution of the concept and hidden variable, i.e., $P(y, H|I, \theta)$. Thus we will seek an approximate computational model for Eq. (4) so that the above computation is tractable and the cost can be reduced.

According to the Bayesian rule and Jensen's inequality, we can factorize the joint distribution in Eq. (4) and find its lower bound,

$$\begin{aligned} \log(P(y|I, \theta)) &= \log \sum_H P(H|I, \theta) P(y|H, \theta) \\ &\geq \sum_H P(H|I, \theta) \log(P(y|H, \theta)) \end{aligned} \quad (5)$$

The sum in the second line in Eq. (5) is the lower bound of Eq. (4). We would rather compute the lower-bound to approximate the log-likelihood in Eq. (4), i.e.,

$$\log(P(y|I, \theta)) \approx \sum_H P(H|I, \theta) \log(P(y|H, \theta)) \quad (6)$$

The two terms in the equation defines the HME based visual concept model. Their definitions are finalized as follows.

The first term in the right hand side is the probability of one mapping configuration given the observed visual features. It depends on the visual features. The second term explains how much a visual concept y is generated for a fixed configuration. When the two terms are known, the concept model is determined.

The first term bridges the concept model with the low-level patch features. For simplicity, here the patches are assumed to be independent and the spatial dependency is not considered. Thus the visual part identity of each patch will be predicted by its observed patch feature. The probability of a mapping configuration can be factorized as,

$$P(H|I, \theta) = \prod_i P(h(i)|x(i), \theta) \quad (7)$$

where $P(h(i)|x(i), \theta)$ is a probability measure of the i -th patch belonging to $h(i)$ -th visual part. In the paper, the visual part is modeled by the Gaussian distribution. For a visual concept y with K parts, the visual part models are denoted as, $P_j^y = N(x|\mu_j^y, \Sigma_j^y)$, $j \in [1, K]$, $y \in [1, M]$. The probability of the i -th patch assigned to the j -th part is approximated as,

$$P(h(i) = j|x(i)) = \frac{1}{Z(x(i))} \left(P_j^y \cdot N(x(i)|\mu_j^y, \Sigma_j^y) \right)^\eta \quad (8),$$

where $Z(x(i)) = \sum_k \left(P_k^y \cdot N(x(i)|\mu_k^y, \Sigma_k^y) \right)^\eta$, P_k^y is a prior probability of the j -th part for the concept y , and η is a smoothing constant.

Up to now, the first part of the definition has been completed. Now we will introduce the definition of the second part. There are many ways to model the conditional probability for a fixed mapping configuration. Here the ME model (see Section 2) is applied to account for the distribution of the visual part configuration. For a possible mapping, the feature extractors (see Eq. (3)) are used to extract the feature and calculate the conditional probability.

Therefore, the complete HME concept model has the parameters for the generative part model, i.e., $P_k^y, \mu_j^y, \Sigma_j^y$, and the weights of feature extractors in the ME, i.e., λ_n . They will be learned through maximizing the log-likelihood on the training set. Now, the objective function is defined as,

$$\Gamma(\theta|T) = \sum_{I,y} \tilde{P}(I,y) \sum_H P(H|I, \theta) \sum_i \lambda_i f_i(H,y) \quad (9),$$

$$- \sum_i \tilde{P}(I) \sum_H P(H|I, \theta) \log Z(H, \theta)$$

where $\tilde{P}(I,y)$ and $\tilde{P}(I)$ are the empirical distribution on the training set.

However, directly optimizing Eq. (9) is still challenging due to the non-linear term, $Z(H, \theta)$. Its lower bound is further exploited for approximation. Its form is,

$$Z(H, \theta) \approx \sum_y \sum_i \frac{f_i(H,y)}{f} \exp(\lambda_i \cdot f) \quad (10)$$

where $f = \sum_i f_i(H,y)$.

3.2 Estimating Model Parameters

The EM-IIS algorithm is developed for optimizing Eq. (9) after substituting Eq. (10) into it. In the E-step, the ME parameters are fixed and we maximize the objective function to get the parameters of visual part models, i.e., P_k^y, μ_j^y and Σ_j^y . That is done using the gradient descent algorithms. Then Eq. (7) is

calculated. In the M-step, the IIS algorithm is used to learn the ME parameters, λ_n while others are fixed. Due to the above approximation, the ME parameters have the closed form (proof is skipped due to the limited space). Figure 1 outlines the EM-IIS algorithm.

-
1. Initialization
 - a) k-means clustering for initializing part weights, means and covariance matrix of part models
 - b) Weights of feature extractors are set to zeros.
 2. M-step: IIS algorithm to update the weights based on the current estimation of part models
 3. E-step: Update parameters of part models using the gradient descent algorithm.
 4. Stop until the predefined criterion is reached, i.e., the maximal iterative number or the relative increment of objective function value is less than a threshold. Otherwise, go to (2).
-

Figure 1 The EM-IIS algorithm for estimating HME model

4. EXPERIMENTS

The HME model can be applied to multi-category image classification problem. However, we will study its capacity on the binary classification problem in the paper, i.e., training the concept model to discriminate an image containing the concept from another without the presence of the concept, i.e., the negative. As discussed above, the ME model is a special case of HME model when the HME concept models share the visual parts and the correspondence is deterministic. Thus, the ME based concept model is trained as the baseline.

4.1 Experimental Setup

The Corel CD image dataset, a popular set for automatic image annotation and retrieval, is used [8]. It has 374 concepts with a total of 5,000 images, 4,500 images for training and 500 for testing. But there are only 260 concepts which have at least one sample in both the training and test sets. Thus our experiments are based on the 260 concepts. For each concept, one concept model and one negative model are trained. To avoid a model becoming bias to the negative class because it has larger training samples than the samples of the concept, we randomly sample equal number of negative samples when training the concept model and the negative model. The detection performances are measured by EER (i.e., equal error rate) and AUC (i.e., area under the ROC curve). The latter is a one-scale quantity of the ROC curve. It is equal to the probability of correctly ranked positive-negative sample pairs in the evaluation set [14].

The SIFT detector is used to extract a set of patches, each being a 128-dimensional appearance feature vector [3]. For the bag-of-words based ME model, four visual terms are learned using the k-means clustering. As a fair comparison, the HME model also uses four visual parts shared by both the concept and the negative models. This setting ensures that the two systems have equal number of parameters. Better results should be observed when more visual parts are used, however, it is not studied here. η is set to be 0.2 empirically.

4.2 Performance Comparison

The HME based system gives an average AUC value of 0.673 over 260 concepts. Compared with 0.610 for the ME based system, the relative improvement is 10.3%. Further concept-by-concept analysis shows that the HME model improves the AUC on 211 concepts, degrades the AUC on 46

concepts and the AUC values have no change on 3 concepts. When comparing the EER, we observe that the HME system has the average EER, 0.354 versus 0.412 for the ME system i.e. a 14.1% relative reduction in average EER. Measured by EER, there are 179 concepts whose EERs are reduced by the HME system, 49 concepts where EERs are increased and 32 concepts where EERs have no change. Therefore, the HME system obviously outperforms the ME system

Now we plot an overall ROC curve across 260 concepts for comparing the HME system with the ME system. To do it, we collect the output scores from 260 concept detectors for the positive samples as the overall positive scores and those scores of all negative samples as the overall negative scores. The scores are not normalized, although normalization may be better. Then the overall ROC curves are drawn in Figure 2 according to the overall positive and negative scores. It shows that the ROC curve of the HME system (red solid curve) has clear advantage over that of the ME system (blue dashed curve).

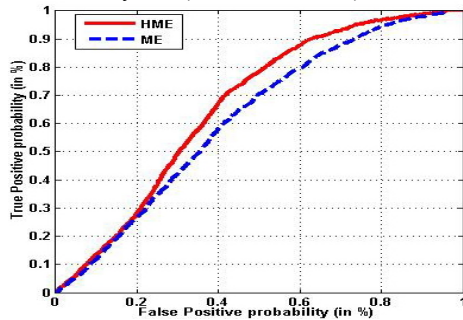
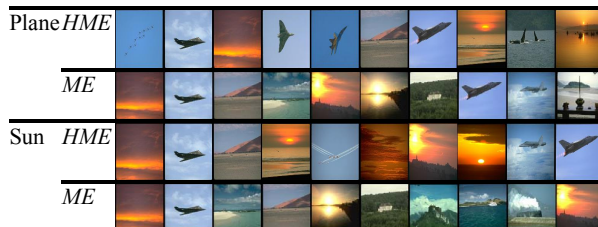
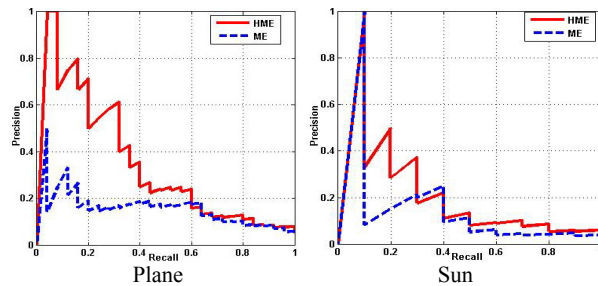


Figure 2 The overall ROC curve comparison between the HME and ME based systems (X-axis: false positive rate. Y-axis: true positive rate. HME: red solid curve. ME: blue dashed curve.)



(a) Top-10 images for *plane* and *sun* using HME and ME



(b) ROC curves for *plane* and *sun* trained with HME and ME

Figure 3 Illustration for visual concepts *plane* and *sun* (a) the top-10 retrieved images (Row HME: top-10 images for the HME system. Row ME: top-10 images for the ME system.) and (b) the corresponding PR-curves (HME: red solid curve. ME: blue dashed curve.)

4.3 Examples of Retrieval

We will now list the retrieval results for 2 selected concepts, i.e.,

plane and *sun*. For each concept, the 500 test images are ranked from the highest score to the lowest according to the log-likelihood ratio between the concept model and its negative model. The top-10 images are depicted in Figure 3a. For each concept, the first row is the result for the HME model and the second is for the ME model. To have an overview of the retrieval performance, the corresponding precision-recall (PR) curves are also depicted in Figure 3b for the two concepts. For the two selected concepts, the precision at the top-10 images for HME model is better than that for the ME model. From the PR curve comparison, the improvement by the HME model is obviously seen at most of the precision-recall points.

5. CONCLUSION

The paper presents a novel hidden maximum entropy algorithm for modeling visual concepts. The HME model combines the good properties of both the bag-of-words approach and the generative model and overcomes each individual drawback. HME allows us to model the local spatial dependency represented in the generative model as well as the image-level statistics captured by the ME model. Our preliminary experiments are carried out on the visual concept detection problem for 260 visual concepts labeled in the Corel dataset. We observe that (1) the relative increment of the average AUC value of 10.3% is obtained when comparing the HME based system with the ME based system; (2) the HME system reduces the average equal error rate from 0.412 to 0.354. In future, we will consider: (1) embedding the shape model and local spatial dependency, (2) evaluating on the multi-category concept recognition and annotation, (3) inferring the hidden correspondence to study its property on segmentation etc.

6. REFERENCES

- [1] A. Bosch, et al., "Scene classification via pLSA", *Proc. of ECCV'06*.
- [2] B.Y. Ricardo & R.N. Berthier, *Modern Information Retrieval*, Addison Wesley, 1999.
- [3] D.G. Lowe, "Object recognition from local scale-invariant features", *Proc. of ICCV'99*.
- [4] G. Csurka, et al., "Visual categorization with bags of keypoints. *Prof. of SLCV Workshop, ECCV'04*.
- [5] J. Jeon & R. Manmatha, "Using maximum entropy for automatic image annotation," *Proc. of CIVR'04*.
- [6] J. Jeon, et al., "Automatic image annotation and retrieval using cross-media relevance models," *ACM SIGIR'03*.
- [7] K. Barnard, et al., "Matching words and pictures," *Journal of Machine Learning Research*, pp. 1107-1135, Vol. 3, 2003.
- [8] P. Duygulu, et al., "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," *Proc. of ECCV'02*.
- [9] R. Fergus, et al., "Object class recognition by unsupervised scale-invariant learning", *Proc. of CVPR'03*.
- [10] S. Gao, et al., "Automatic image annotation through multi-topic text categorization", *Proc. of ICASSP'06*.
- [11] S. Lazebnik, et al., "A maximum entropy framework for part-based texture and object Recognition", *Proc. of ICCV'05*.
- [12] S. Wang, et al., "The latent maximum entropy principle", *Proc. of ISIT'02*.
- [13] A. Berger, et al., "A maximum entropy approach to natural language processing", *Computational Linguistics*, (22-1), March 1996.
- [14] S. Gao, et al., "An ensemble classifier learning approach to ROC optimization", *Proc of ICPR'06*.