

PROPAGATING IMAGE-LEVEL PART STATISTICS TO ENHANCE OBJECT DETECTION

Sheng Gao, Joo-Hwee Lim and Qibin Sun

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

ABSTRACT

The bag-of-words approach has become increasingly attractive in the fields of object category recognition and scene classification, witnessed by some successful applications [5, 7, 11]. Its basic idea is to quantize an image using visual terms and exploit the image-level statistics for classification. However, the previous work still lacks the capability of modeling the spatial dependency and the correspondence between patches and object parts. Moreover, quantization always deteriorates the descriptive power of the patch feature. This paper proposes the hidden maximum entropy (HME) approach for modeling the object category. Each object is modeled by the parts, each having a Gaussian distribution. The spatial dependency and image-level statistics of parts are modeled through the maximum entropy approach. The model is learned by an EM-IIS (Expectation maximum embedded with improved iterative scaling) algorithm. Our experiments on the Caltech 101 dataset show that the relative reduction of equal error rate of 23.5% and relative improvement of AUC (area under ROC) of 22.0% are obtained when comparing the HME based system with the ME based baseline system.

Index Terms – object detection, hidden maximum entropy, EM-IIS, bag-of-words, area under ROC.

1. INTRODUCTION

The bag-of-words approach is commonly adopted in text information retrieval: a text document is represented using the word occurrence [3]. Although the method ignores syntactic or semantic information, its success has been proven in text categorization and retrieval. Recently, the bag-of-words has attracted attention in the community of image related pattern recognition such as scene classification [2], object recognition (e.g. [5, 7, 11]), etc. However, unlike text document, image document is not symbolic. The clustering algorithms (e.g. k-means) have to be employed to quantize the patch features into a finite set of visual words. A high-dimensional feature vector is then extracted. Finally, machine learning algorithms such as support vector machine [5], translation model [7], maximum entropy (ME) [1], MC MFoM [10], probabilistic latent semantic analysis [2], are applied for classification.

The bag-of-words could easily make use of image-level statistics. The quantities, e.g., unigram and bigram of visual terms, are invariant to rotation and are robust to partially occlusion. This is the main reason for its success in object recognition. However, it still lacks the capability of modeling the spatial dependency and the correspondence between patches and object parts which are demanded in some applications (e.g. object localization). In addition, quantization also results in the loss of some discriminative information in continuous visual features. Such loss is not recoverable regardless of how powerful the object model is.

To discover the correspondence between the patches and object parts, the generative object model has been developed [6, 8, 12], where each object has a few parts and each part is modeled by a Gaussian distribution. The correspondence is hidden and is learned using the EM algorithm. Such modeling is feasible to integrate the appearance, shape and local spatial dependency. But it also fails in capturing the image-level statistics of parts.

In the paper, the hidden maximum entropy (HME) approach is presented for modeling object categories aiming to obtain the joint benefits from both the bag-of-words approach and generative model. The HME model has some parts, each being a Gaussian distribution, which connect the patch feature with the discrete symbols, i.e., parts. The part configuration in the image and their interaction are also modeled through the ME model. Since the correspondence is hidden, a feasible EM-IIS algorithm, i.e., EM embedded with improved iterative scaling (IIS), is developed for learning. Note that the HME directly characterizes the distributions of the patch feature and part configuration. Therefore quantization is not necessary. When the correspondence is deterministic and the object categories share the visual parts, the HME would become the ME. Section 3 will detail more on it.

The paper is organized as the following. In the next section, the ME is introduced. Then the HME model is presented. The experiments are reported in Section 4. Finally, conclusion is made in Section 5.

2. MAXIMUM ENTROPY MODEL

The maximum entropy model has been successfully applied to the text document [1]. However, the natural image representation consists of a set of patch features (e.g., color,

texture, etc.). Thus, we have to “convert” image into a text-like (i.e., symbolic) document. The first step is to learn visual terms using the clustering algorithms such as k-means. Then a patch is assigned to its closest term based on its distance with the centers of visual terms. Quantization finally makes an image document to be a symbolic document so that any technique (including feature extraction and modeling algorithms) developed for text categorization and retrieval can be applied. Among all proposed approaches, the ME model is promising because of its capability to fuse diverse features.

Assuming there are M object categories and a training set, $T = \{(I_i, y_i)\}$, where (I_i, y_i) is an image I_i and its label, y_i , then M object models should be trained for M -class classification. We denote the extracted visual features as, $I_i = (x_i(1), x_i(2), \dots, x_i(L_i))$, where $x_i(\cdot)$ is a D -dimensional vector and L_i is 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 $x_i(\cdot)$.

2.1. Maximum Entropy Model

To make a decision, the predicted probability of an object, $P(y|I, \theta)$, is calculated, where y is the object category, I is the symbolized image and θ is the model parameters. Then the image is assigned to the object, y^* , which gives the maximal predicted probability, i.e.,

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

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 describe that the empirical quantity of each feature must be equal to its predicted value by the learned model. This results in the following form of the object category model (Refer to [1] for details).

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

where $Z(I, \theta) = \sum_y \exp(\sum_i \lambda_i \cdot f_i(I, y))$, $\theta = \{\lambda_1, \lambda_2, \dots, \lambda_N\}$. In the paper, the unigram feature is used,

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

where q is a visual term, c is an object category 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 object model in Eq. (2) has its own distinct parameters.

The model parameters can be obtained by maximizing the log-likelihood in the training set. Usually the generalized

iterative scaling (GIS) or IIS algorithm is employed for efficient estimation [1].

2.2. Limitation Analysis of ME

With carefully designed feature extractors, the ME based object model could capture the image-level statistics of visual terms which are invariant to rotation and are robust to partially occlusion. However, the spatial dependency is not characterized. It cannot answer the correspondence due to the image-level representation. Moreover, unavoidable quantization error cannot be recovered by the following process. Further, 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 therefore challenging to incorporate the shape model because the shape model needs the spatial configuration of object parts.

In next section, we will introduce the hidden maximum entropy approach to address all these issues. The proposed method takes learning visual terms and estimating the correspondence as a hidden stage. The HME model then jointly characterizes the distribution of patch features and part configuration.

3. HIDDEN MAXIMUM ENTROPY BASED OBJECT CATEGORY MODELING

The bag-of-words based image concept models consist of a universal visual term models to map the patches in the image to their closest visual terms as well as a concept dependent model to characterize the image-level statistics of the symbolized image. In the HME model, each object has some parts, which generate the observed visual patch feature, and the distribution of the part configuration is characterized by the ME model. The correspondence between the object parts and the image patches are the hidden random variable.

3.1. Hidden Maximum Entropy

For the image patch representation, $I = (x(1), x(2), \dots, x(L))$, and the object category model, there is a hidden random variable, $H = (h(1), h(2), \dots, h(L))$, to describe the mapping between the object parts and the patches. Here $h(i)$ is one of K parts. Thus the log-likelihood for predicting the object category 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 object model and L -patches image, there will be K^L correspondences. Thus, the computation cost for the above sum calculation is very huge. Even it is possible, there is another challenge to find a computable function for the joint distribution of object category and hidden variable, i.e., $P(y, H|I, \theta)$. So we have to seek an approximate computational model so that 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 object category model. In the next we will finish their definitions.

The first term in the RHS is the probability of one mapping configuration given the observed visual features. It is computed from the visual features. The second term explains how much the object category y is generated for a fixed configuration. When the two terms are given, the object category model is determined.

The first term bridges the object model with the low-level patch features. For simplicity, here the patches are assumed independent and the spatial dependency is not considered. Thus the identity of each patch will be predicted by itself 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 part. In the paper, a Gaussian distribution is used to model the object part. For an object category y with K parts, the 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 object y , and η is a smoothing constant.

Up to now, the first part definition has been defined. Now we will complete 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 features in the mapping. For a possible mapping, the feature detectors (see Eq. (3)) are used to extract feature and calculate the conditional probability.

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

$$\begin{aligned} \Gamma(\theta|T) &= \sum_{I,y} \tilde{P}(I,y) \sum_H P(H|I, \theta) \sum_i \lambda_i f_i(H,y) \\ &\quad - \sum_i \tilde{P}(I) \sum_H P(H|I, \theta) \log Z(H, \theta) \end{aligned} \quad (9)$$

where $\tilde{P}(I,y)$ and $\tilde{P}(I)$ are the empirical distributions.

3.2. Estimating Model Parameters

The feature extractors used in the ME are a linear function of the part occurrences for a fixed mapping and it can be summed over all patches. Then the sum in the first line in Eq. (9) can be efficiently computed due to the independence of patches. However, the sum in the second line is still difficult due to the non-linear term, $\log Z(H, \theta)$. We will further approximate Eq. (9) for efficient computation. It is noted that

$$-\log Z(H, \theta) \geq 1 - Z(H, \theta) \quad (10)$$

and,

$$Z(H, \theta) \leq \sum_y \sum_i \frac{f_i(H,y)}{f} \exp(\lambda_i \cdot f) \quad (11)$$

where $f = \sum_i f_i(H,y)$. It is a constant here.

Substituting Eqs. (10-11) into Eq. (9) will result in a new lower bound, which is a linear function over all parts and patches. The EM-IIS algorithm is used for learning the parameters. In the E-step, the ME parameters are fixed and we maximize the objective function to estimate the parameters of the part models, i.e., P_k^y, μ_j^y and Σ_j^y . In the M-step, the IIS algorithm is used to learn the weights of feature extractors, λ_n , while others are fixed. Because of the normalized feature extractors in Eq. (3), a closed solution can be found. The proof is skipped due to the limited space.

4. EXPERIMENTAL RESULTS AND ANALYSIS

The HME based object model can be applied to multi-category classification. In the paper we will demonstrate its power on the object detection. We will train the HME based object models to discriminate the image containing the object from the background image. Because of its relation with ME, the ME model is trained as the benchmark [11].

4.2. Experimental Setup

The Caltech-101 database is used for evaluation, which contains 101 object categories plus 1 background category. There are totally 9,144 images, ~ 40 to 800 images per category and ~ 50 for most categories [12]. We select 15 images, whose numbers are from 1 to 15, per category for training without any intention. Thus 1,530 images are reserved for training and all the left are used for testing. All images are resized so that its maximal dimension is not more than 300 pixels. The SIFT detector is used to extract a set of patches, each being a 128-dimensional appearance feature vector [4].

For the bag-of-words based ME model, 4 visual terms are learned using the k-means clustering. As a fair comparison, the HME model also has 4 visual parts shared

by the object and the background model. The setting makes the two systems have the same size of parameters. Better results should be gotten with more parts, however, it is not our intention here.

4.3. Experimental Results

The two metrics are used to evaluate the detection, i.e., the average equal error rate (EER) over the 101 categories and average AUC (area under the ROC. see [9]). The latter is a one-scale quantity of ROC curve and measures how much probability the positive-negative image pair are correctly ranked.

The average EER is 33.6% for the HME models while it is 43.9% for the ME. Thus HME gives rise to 23.5% reduction of the average EER. The concept-by-concept analysis shows that HME improves the detection performance among 87 categories out of the 101 categories, there are 11 categories whose performances become worse, and others have no change. When measured with the AUC metric, we observe that (1) HME increases the average AUC, 58.2%, for ME to 71.0% and (2) HME is better than ME among 96 categories, 4 categories becoming worse. Therefore, the HME model outperforms ME significantly.

| Top-10 representative patches | Weights (Object/Background) | Object |
|-------------------------------|--------------------------------|-----------------|
| | -0.58/-0.82 | Car_side |
| | -0.94/-0.50 | |
| | -1.4/-0.28 | |
| | -0.38/-1.15 | |
| | -0.46/0.43 | ME |
| | -0.73/0.66 | |
| | -0.56/0.51 | |
| | 2.00/-2.86 | |
| | -1.02/-0.45 | Schooner |
| | -1.53/-0.25 | |
| | -0.31/-1.32 | |
| | -0.33/-1.28 | |
| | -2.33/1.56 | ME |
| | -1.92/1.25 | |
| | 1.70/-3.08 | |
| | 1.71/-2.88 | |

Figure 1 Top-10 representative patches of part models and weights of the object (first value) and background models (second value) for the two selected categories *car_side* and *schooner*.

4.3. Illustration of Learned Visual Parts

Now we visualize the part models learned from HME and k-means clustering for two selected object categories. One is *car_side* on which HME is better than ME (EER: 38.0% vs. 46.4%) and another is *schooner* on which HME is worse (ERR: 25.0% vs. 22.1%). For each part model, the top-10 training patches closest to the part center are chosen as its representative and are cropped from the corresponding image based on its location and scale. Then they are resized to 11x11 pixels.

The top representative patches are depicted in Figure 1. The first column is the image listing representative patches.

Each image has 4 rows corresponding to 4 part models for each object category and model. The second column is the learned weights of feature extractors in ME for the object and background models (see Eq. (2)), the first value for the object and the second for the background. The value measures the importance degree of part to the model. The part having higher value is more representative than that with lower value. For example, for *car_side* model learned with HME, the forth and first parts are representative patterns of *car_side* while the other two are background patterns. These images clearly show the distinct patterns of the object and background. The background patterns learned with HME is more consistent than ME, i.e., it has fewer object patterns in its top-10, even for *schooner* on which ME operates better.

5. CONCLUSION

This paper presents the HME-based object category model to integrate the good properties of ME model and the generative model. It captures not only the image-level statistics of the object parts but also the local characteristics of the patches through the generative model. The proposed method is evaluated on the object detection for 101 object categories. Compared with the ME, 23.5% relative reduction of EER is obtained and the AUC relative improvement reaches 22.0%. In future, we shall study the effect of class-specific parts (here they are shared) and segmentation based on the mapping, and the improved algorithm for reducing computation.

6. REFERENCES

- [1] A. Berger, et al., "A maximum entropy approach to natural language processing", Computational Linguistics, (22-1), 1996.
- [2] A. Bosch, et al., "Scene classification via pLSA", Proc. of ECCV'06.
- [3] B.Y. Ricardo & R.N. Berthier, Modern Information Retrieval, Addison Wesley, 1999.
- [4] D.G. Lowe, "Object recognition from local scale-invariant features", Proc. of ICCV'99.
- [5] G. Csurka, et al., "Visual categorization with bags of keypoints. Prof. of SLCV Workshop, ECCV'04.
- [6] M. Weber, et al., "Unsupervised learning of models for recognition", Proc of ECCV'00.
- [7] P. Duygulu, et al., "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," Proc. of ECCV'02.
- [8] R. Fergus, et al., "Object class recognition by unsupervised scale-invariant learning", Proc. of CVPR'03.
- [9] S. Gao, et al., "An ensemble classifier learning approach to ROC optimization", Proc of ICPR'06.
- [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] L. Fei-Fei, et al., "One-shot learning of object categories", IEEE Trans. on PAMI, Vol.28, No.4., pp.594 – 611, 2006.