LATENT LAYOUT ANALYSIS FOR DISCOVERING OBJECTS IN IMAGES

David Liu[1]*, Datong Chen*[2]*, and Tsuhan Chen*[1]

Department of Electrical and Computer Engineering[1], School of Computer Science[2]
Carnegie Mellon University, Pittsburgh, U.S.A.
dliu@cmu.edu[1], datong@cs.cmu.edu[2], tsuhan@cmu.edu[1]

ABSTRACT

Latent Layout Analysis (LLA) is a novel unsupervised learning technique to discover objects in unseen images using a set of un-annotated training images. LLA defines a generative model that associates latent aspects to local appearances. The dependency between aspects and position is captured by a spatial sensitive aspect model. This dependency distinguishes LLA from Probabilistic Latent Semantic Analysis (PLSA). The latent aspects together with the latent layout constitute a compact scene representation. We demonstrate that the proposed LLA significantly outperforms Probabilistic Latent Semantic Analysis in two tasks: object discovery (detection) and object localization.

1. INTRODUCTION

Unsupervised image understanding systems have many advantages comparing to supervised systems due to the difficulty of image annotation. First, an image may consist of many objects in a complex layout. So far there is no common approach to annotating images at the object level. Second, there are many visual illusions showing that different people may have different understandings of an image. Third, object level annotation in a supervised system requires manually labeling the location and categorization of each object in images, which is very time consuming. It is very expensive to collect large mount of accurate annotated images for constructing a supervised image understanding system. On the contrary, training an unsupervised system does not need annotated images. Considering the abundance of images available on the Internet, unsupervised learning methods provide a promising direction.

One unsupervised learning method called Probabilistic Latent Semantic Analysis (PLSA) [1] has recently been applied to the image understanding domain [2][3] and shown to outperform classical clustering methods such as k-means. This model has earlier been used in the text and linguistic domains. PLSA is a generative model, and can be used to interpret how a document of words is generated. A document is considered as a mixture of "aspects" (or "topics"), and each aspect consists of a mixture of words. The power of PLSA originates from the fact that aspects can be learned in an unsupervised manner given a set of document-word pairs. In the image understanding domain, documents are analogous to images, and words are analogous to visual words. Applying PLSA to pairs of image-visual words is hence also capable of extracting latent aspects.

Fig. 1. Probabilistic Latent Semantic Analysis ignores the spatial layout of visual words in an image.

One important drawback of PLSA, however, is that the set of document-word pairs ignores the geometric layout of words in an image; this is illustrated in Fig. 1. PLSA as a generative model is hence specifying an image without spatial structure. In other words, if we arbitrarily shuffle the local features in the image around, we get the same latent aspects! As a result, the performance of PLSA still leaves room for improvement.

Here we propose Latent Layout Analysis (LLA) which explicitly considers spatial structure. LLA associates latent aspects to spatial structure, in the sense that latent aspects become position-dependent. We will describe LLA in detail in Section 4. Before that, we will describe how to generate visual words in the next section, and briefly review PLSA in Section 3. Finally, in Section 5, we will show LLA performs significantly better than PLSA.

2. GENERATING VISUAL WORDS

Visual words are the basic units that form the observations of an image. The system needs to determine a number of regions to generate the visual words from. These regions are determined by running the Canny edge detector and then uniformly sampling points from all edges. These points are called interest points. Scale Invariant Feature Transform (SIFT) image

features [4] around the interest points are computed. This procedure [5] provides a set of local feature vectors. Note that SIFT image features are general and can be applied to a wide range of different objects and tasks [4]. To obtain a finite set of visual words, we perform k-means clustering on all local feature vectors from all training images. The resulting cluster centers form the dictionary of visual words $\{w_1, ..., w_W\}$. For each training or test image, its visual words are obtained by choosing the closest w_i for each of its local feature vectors.

3. PROBABILISTIC LATENT SEMANTIC ANALYSIS

PLSA is a generative model which describes the process an image is formulated as in Fig. 2. Suppose the training data consists of D images (or documents in the text and linguistic domains), $\{d_1, ..., d_D\}$. Each image is considered as a mixture of aspects: $P(z_k|d_i)$ is the probability of aspect z_k occurring in image d_i. Assume there are a predefined number of Z latent aspects, $\{z_1, ..., z_Z\}$. Using inference methods, it is then possible to infer latent variables of the generative model. Each aspect is further considered as a mixture of words: $P(w_j|z_k)$ is the probability of word w_j occurring in aspect z_k. We denote W as the total number of (visual) words, $\{w_1, ..., w_W\}$. The joint p.d.f. of document, aspects and words is therefore formulated as

$$P(d_i, w_j, z_k) = P(w_j|z_k)P(z_k|d_i)P(d_i). \quad (1)$$

The prior probability $P(d_i)$ is modelled as a multi-nomial distribution of words: $P(d_i) \propto \sum_j n(d_i, w_j)$, where $n(d, w)$ is the image-word co-occurrence table, and $n(d_i, w_j)$ denotes the number of occurrences of w_j in d_i.

To estimate the distribution of latent variables $P(w_j|z_k)$ and $P(z_k|d_i)$ that maximize the likelihood of the joint probability, PLSA model employs the standard Expectation-Maximization (EM) algorithm [6]. The EM algorithm consists of two steps: the E-step computes the posterior probabilities for the latent variables; the M-step maximizes the expected complete data likelihood. Derivations of the E- and M-step of PLSA can be found in [1]. Here we restate the results:

E-step:

$$P(z_k|d_i, w_j) \propto P(w_j|z_k)P(z_k|d_i) \quad (2a)$$

M-step:

$$P(w_j|z_k) \propto \sum_i n(d_i, w_j)P(z_k|d_i, w_j) \quad (2b)$$

$$P(z_k|d_i) \propto \sum_j n(d_i, w_j)P(z_k|d_i, w_j) \quad (2c)$$

$$P(d_i) \propto \sum_j n(d_i, w_j) \quad (2d)$$

Note that these equations need normalization to make them probability distributions. In summary, given $n(d_i, w_j)$, maximum likelihood fitting by the EM algorithm yields $P(w_j|z_k)$, $P(z_k|d_i)$, and $P(d_i)$.

During inference, given a test image d_{query}, the factors $P(z_k|d_{\text{query}})$ are computed using the "fold-in" technique described in [1]; the EM algorithm is run in the same way as in learning, but now keeping the factors $P(w_j|z_k)$ obtained in the learning stage fixed.

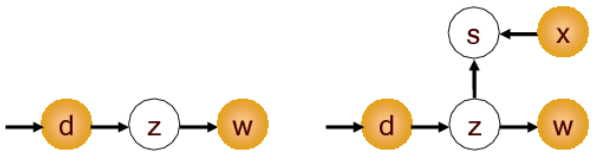

Fig. 2. The probabilistic graphical models, PLSA (left) and LLA (right). Observed variables are marked by filled yellow. The direction of arrows indicate the Bayesian dependencies.

4. LATENT LAYOUT ANALYSIS

To exploit spatial structure among aspects, we propose the Latent Layout Analysis (LLA) model as shown in Fig. 2. In LLA, in addition to the latent aspects, there exist a number of S latent layout variables $\{s_1, ..., s_S\}$ and the spatial locations of visual words $\{x_1, ..., x_P\}$.

The spatial location dictionary x is obtained by sampling P points uniformly from the 2-D image coordinates. Each visual word w_j is then associated with the spatially closest position $x_p \in \{x_1, ..., x_P\}$. Therefore, an image d_i is not a bag-of-word but a bag-of-word-position, denoted by d_i^s. The joint probability of LLA is then computed as:

$$\begin{aligned} &P(d_i^s, s_l, z_k, x_p, w_j) \\ &= P(w_j|z_k)P(s_l|z_k, x_p)P(z_k|d_i^s)P(d_i^s)P(x_p) \end{aligned} \quad (3)$$

Let us define the image-word-position co-occurrence table $n(d, w, x)$, with $n(d_i^s, w_j, x_p)$ denoting the number of occurrences of word w_j at position x_p in image d_i^s, replacing the image-word co-occurrence table, $n(d, w)$, in PLSA.

LLA models each image as a mixture of aspects, each associated with a spatial layout, $s_l \in \{s_1, ..., s_S\}$. Intuitively, a layout encodes the rough locations of foreground objects. Introducing layout s_l allows LLA to exploit the spatial layout information of aspects and provides a better interpretation of the image than PLSA does. In the next section, we introduce a model that relates aspects to layouts.

4.1. Spatial sensitive aspect model

To simplify our discussion, here we focus on the case $z_k \in \{z_1, z_2\}$, where we interpret the latent aspect z_1 (z_2) as the presence (absence) of a foreground object in the image. We call z_1 the foreground aspect and z_2 the background aspect.

Aspects are layout-dependent, which is one of the major contributions of LLA. Notice that if aspects were merely dependent on position, then the flexibility of the system would still be confined, because it cannot model the case where different images have different aspect distributions at the same position. LLA is able to handle this case by a spatial sensitive aspect model, $P(z_k|s_l, x_p)$. We model the spatial sensitive aspect model of the foreground aspect z_1 by

$$P(z_1|s_l, x_p) = exp(\frac{-(x_p - \mu_l)^2}{2\sigma_l^2}) \quad (4)$$

for $s_l \in \{s_1, ..., s_{S-1}\}$, where μ_l and σ_l are control parameters associated with layout s_l. We will explain in Section 5 how to obtain these parameters unsupervised. The background aspect distribution is then $P(z_2|s_l, x_p) = 1 - P(z_1|\ s_l, x_p)$. Eq.(4) says that, given the spatial layout $s_l \in \{s_1, ..., s_{S-1}\}$, the closer the position of the visual word is to the peak of the foreground aspect spatial distribution, the more likely a foreground object is located at that position. Since an image consists of a mixture of spatial layouts s_l, each associated with a single-mode distribution, the system is hence capable of modelling complex spatial distributions in the scene.

To handle the case where no foreground object is present in the scene, we define a layout s_S for an image consisted of only background clutter and no foreground object. For $s_l = s_S$, the spatial sensitive aspect model is defined as a uniform distribution, indicating the conditional independency of x_p and z_1.

4.2. Model fitting with the EM algorithm

The goal is to maximize the log-likelihood,

$$\mathcal{L} = \sum_i \sum_j \sum_p n(d_i^s, w_j, x_p) \log P(d_i^s, w_j, x_p) \quad (5)$$

This can be achieved by the EM algorithm. Here we state the results:

E-step:

$$P(s_l, z_k|d_i^s, w_j, x_p) \propto P(z_k|d_i^s)P(w_j|z_k)P(s_l|z_k, x_p) \quad (6)$$

M-step:

$$P(w_j|z_k) \propto \sum_i \sum_p n_{ijp} \sum_l P(s_l, z_k|d_i^s, w_j, x_p) \quad (7)$$

$$P(z_k|d_i^s) \propto \sum_j \sum_p n_{ijp} \sum_l P(s_l, z_k|d_i^s, w_j, x_p) \quad (8)$$

$$P(d_i^s) \propto \sum_j \sum_p n_{ijp} \quad (9)$$

$$P(x_p) \propto \sum_i \sum_j n_{ijp} \quad (10)$$

where $n_{ijp} \equiv n(d_i^s, w_j, x_p)$. Note that these equations need normalization to make them probability distributions. Since

$$P(s_l|z_k, x_p) \propto P(z_k|s_l, x_p)P(s_l|x_p), \quad (11)$$

where $P(s_l|x_p)$ is assumed uniform for simplicity, we can embed the spatial sensitive aspect model into the E-step.

Both learning and inference use the above maximum likelihood fitting to obtain the conditional probabilities, except that during inference the factor $P(w_j|z_k)$ is kept fixed, which is called the 'fold-in' method in [1].

5. EXPERIMENTS

In our experiments, we use 100 face images and 100 non-face images and set the number of latent aspects to two. Instead of using the common terminology of face "detection" in supervised systems, it is more appropriate to say the system "discovers" objects since the system is unsupervised. Note that the visual words are neither obtained from labelled data, nor are they specifically designed for this face/non-face task, implying generality for other tasks and the unsupervised nature of this system.

The PLSA code is made online by [5], and we use the same set of images as [5], where the face images are taken from the first 100 images in the Caltech face dataset [7]. As in [5], images are resized to around 200×140, and color information is discarded.

Eq.(4) involves μ_j and σ_j. We obtain them by fitting PLSA to the data, after which we obtain $P(z|d, w)$. This can be used as a set of weightings, indicating how likely a visual word belongs to the foreground aspect. We can then compute the weighted mean and weighted variance of the positions of the visual words for each image. By k-means clustering these weighted means of all images, we obtain clusters $\{\mu_1, ..., \mu_{S-1}\}$. We simply use the average of the weighted variance of all images as the variance $\sigma_l = \sigma$. In our experiments, for $S = 4$, we obtain $\mu_1 = (80.1, 68.7)$, $\mu_2 = (102.5, 67.2)$, $\mu_3 = (123.2, 68.3)$, and $\sigma = (45.9, 28.0)$.

Depending on the convergence rate of the EM algorithm, training and testing using LLA both take around 6 seconds per image on a Pentium-4 3.4 GHz machine.

To decide the presence/absence of the foreground object in the scene, in PLSA, we compute $P(d_i|z_1) \propto P(z_1|d_i)P(d_i)$ for each image d_i; the higher it is, the more likely that image contains a foreground object. In LLA, we compute

$$P(d_i^s|z_1, s \neq s_S) = \alpha P(s \neq s_S|z_1)P(z_1|d_i)P(d_i) \quad (12)$$

where

$$P(s \neq s_S|z_1) = \alpha' \sum_{x_p} P(z_1|s \neq s_S, x_p)P(s \neq s_S|x_p)P(x_p) \quad (13)$$

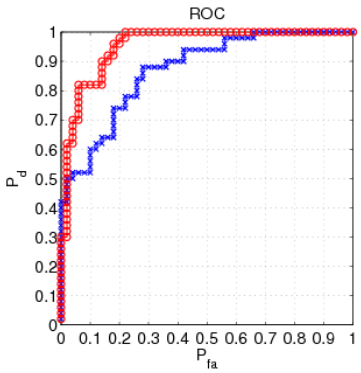

Fig. 3. ROC curve of LLA (red circles) and PLSA (blue crosses).

and α and α' are normalization constants. Fig. 3 shows the ROC curve and it can be seen that LLA performs significantly better than PLSA in this discovery (detection) task.

The most likely latent aspect of each visual word in PLSA can be computed from the posterior in Eq.(2a):

$$z^* = \arg\max_z P(z|d_i, w_j) \tag{14}$$

In LLA, if a foreground object is discovered (based on a threshold of equal error rate), we determine the most likely aspect of each visual word by

$$z^* = \arg\max_z P(z|d_i^s, w_j, x_p, s \neq s_S). \tag{15}$$

Otherwise, we determine the optimal labels by

$$z^* = \arg\max_z P(z|d_i^s, w_j, x_p, s = s_S). \tag{16}$$

The red and green ellipses in Fig. 4 represent the inferred most likely aspects z^* of each visual word; red indicates that the system labels the particular region as face. Comparing PLSA (row 2) to LLA (row 4), it can be seen that faces are more correctly located by LLA. Also, PLSA falsely labels many background non-face objects by red ellipses. In row 3 and row 5, labelling results are expressed in masked images. We superimpose the effects of the red and green ellipses by creating a mask with 1's and -1's on the red and green ellipse centroids, and convolve a Gaussian kernel over the mask. The mask is then applied on the original image. It can be seen that, LLA (row 5) rejects the non-face background objects much better than PLSA (row 3).

6. CONCLUSION AND FUTURE WORK

We proposed a novel technique called LLA to discover objects in unseen images using a set of un-annotated training images. LLA defines a generative model that associates latent aspects to local appearances. The dependency between aspects and position is captured by a spatial sensitive aspect model. Even though only justified on an image understanding task, this spatial dependency also exists in documents, and hence we expect LLA to perform better than PLSA in text understanding tasks as well. Extending the experiments and perhaps the framework to more complicated objects is of future interest.

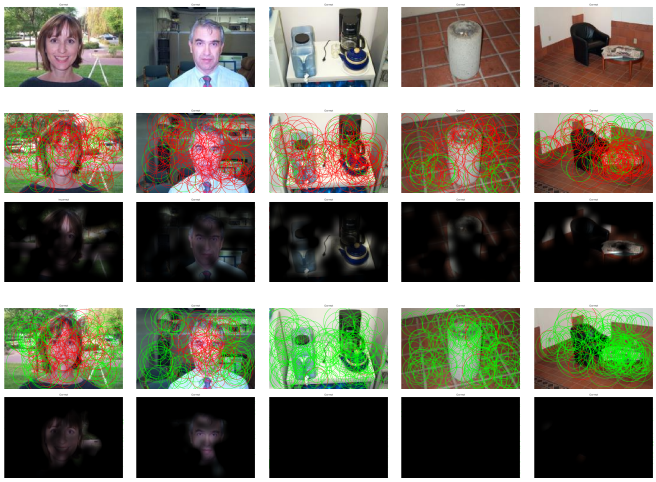

Fig. 4. Second row and third row: result of PLSA. Last two rows: result of LLA. Best viewed in color.

7. ACKNOWLEDGEMENT

We acknowledge Rob Fergus for making the PLSA code online [5]. This work is supported by the Taiwan Merit Scholarship TMS-094-1-A-049 and by the ARDA VACE program.

8. REFERENCES

[1] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, pp. 177–196, 2001.

[2] J. Sivic, B. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman, "Discovering objects and their location in images," in *IEEE Intl. Conf. on Computer Vision (ICCV)*, 2005.

[3] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool, "Modeling scenes with local descriptors and latent aspects," in *IEEE Intl. Conf. on Computer Vision (ICCV)*, 2005.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[5] Rob Fergus, *ICCV short course 2005,* `http://people.csail.mit.edu/fergus/iccv2005/bagwords.html`.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[7] *Caltech face dataset.* `http://www.vision.caltech.edu/Image_Datasets/faces`.