UNSUPERVISED IMAGE LAYOUT EXTRACTION

David Liu, Datong Chen, Tsuhan Chen*

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

ABSTRACT

We propose a novel unsupervised learning algorithm to extract the layout of an image by learning latent object-related aspects. Unlike traditional image segmentation algorithms that segment an image using feature similarity, our method is able to learn high-level object characteristics (aspects) from a large number of unlabelled images containing similar objects to facilitate image segmentation. Our method does not require human to annotate the training set and works without supervision. We use a graphical model to address the learning of aspects and layout extraction together. In particular, aspect-feature dependency from multiple images is learned via the Expectation-Maximization algorithm. We demonstrate that, by associating latent aspects to spatial structure, the proposed method achieves much better layout extraction results than using Probabilistic Latent Semantic Analysis.

1. INTRODUCTION

Automatic image layout extraction provides crucial information for content-based image processing and understanding. An image can be considered as a set of objects arranged in a spatial layout. The layout is one of the basic cues to understand the image and also can be used to guide object analysis. Layout extraction is different than image segmentation even though image segmentation also examines the composition of an image. In comparison, image segmentation methods partition an image into regions that consist of similar color, texture, or position. This partition often operates on a single image without using any knowledge about objects (Section 4 has more discussions). Image layout extraction partitions an image by objects and therefore requires learning of object information from multiple images. A sample result of image segmentation and image layout extraction is illustrated in Fig. 1. It illustrates a common segmentation result obtained by Blobworld segmentation [1] which uses color, texture, and position features, and a common image layout extraction result of our proposed method. Notice the cleaner result obtained in Fig. 1(c) even without using color information; it is a result of being able to learn from multiple images containing similar objects. In this paper, we propose an unsupervised learning approach to extract image layout from a bunch of unlabelled images.

(a) (b) (c)

Fig. 1. (a) Original image (b) Blobworld segmentation (c) layout extraction using latent aspect learning.

Image layout extraction is object-related but is different than two other common image understanding problems: object detection and object localization. Object detection aims at deciding whether or not the image contains the object of interest; object localization finds out the location of the object of interest. In both problems, the object of interest is specified beforehand. In image layout extraction, the object of interest is unknown beforehand; the goal of the system is to understand the composition of the image, and hence it becomes irrelevant what specific objects compose the image.

We propose to learn unspecific object characteristics by an unsupervised approach. Unsupervised image understanding systems have many advantages compared to supervised systems due to the difficulty of image annotation. First, an image may consist of many objects arranged in a complex layout. So far there is no common approach to annotating images at 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 labelling 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.

The method we propose has its root in an unsupervised learning method called Probabilistic Latent Semantic Analysis (PLSA)[2], which has recently been applied to the image understanding domain [3][4]. This model has earlier been

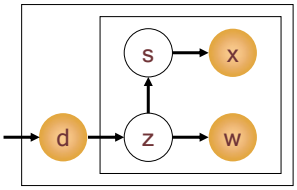

Fig. 2. The proposed probabilistic graphical model.

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" ("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. It is hence capable of accumulating information from multiple documents.

One important drawback of PLSA, however, is that the set of document-word pairs ignores the layout or order of words in a document. PLSA as a generative model is hence generating a document without structure. In other words, if we arbitrarily shuffle the words in the document around, we get the same latent aspects (topics)! As a result, the performance of PLSA still leaves room for improvement. Sivic et al. [3] add spatial information into the feature representation and leave the PLSA model unchanged. Our method extends the PLSA graphical model and explicitly models spatial structure. As in [5], latent aspects are associated with word positions, but in this work we do not require the single-modal Gaussian assumption.

2. STATISTICAL FORMULATION

The algorithm we propose is based on the graphical model shown in Fig. 2. It is a generative model and describes the process of composing an image. This graphical model is trained using a number of un-annotated training images, $\{d_1, ..., d_D\}$. We use the 'plate' notation of a box surrounding nodes to denote that those variables are replicated as many times as their number of realizations. Observed variables are marked by darker nodes.

2.1. Feature descriptor

From the D training (yet un-annotated) images, the system determines a number of regions to generate the observation features. 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 [6] around the interest points are computed. This procedure 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 [6].

To reduce the number of states, we quantize the image features into a finite set of visual words [7]. This is achieved by running k-means clustering on all local feature vectors from all training images. The resulting cluster centers form the dictionary of visual words $\{w_1, ..., w_W\}$. Visual words serve as the basic units that form the observations of an image. For each training or test image, its visual words are obtained by choosing the closest w_i for each of its local feature vectors.

2.2. Generative model

The joint probability of this graphical model is computed as:

$$\begin{aligned} &P(d_i, s_l, z_k, x_p, w_j) \\ &= P(d_i)P(z_k|d_i)P(w_j|z_k)P(s_l|z_k)P(x_p|s_l) \end{aligned} \tag{1}$$

An 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. In the linguistic domain [2], aspects can be considered as topics of a document; in image processing, aspects can be considered as objects that constitute an image. Assume there are a predefined number of Z latent aspects, $\{z_1, ..., z_Z\}$. Using EM algorithm, it is possible to infer $P(z_k|d_i)$, as we will see later.

Each aspect is further considered as a mixture of visual words: $P(w_j|z_k)$ is the probability of visual word w_j occurring in aspect z_k. We denote W as the total number of visual words, $\{w_1, ..., w_W\}$. Intuitively, different objects correspond to different features, and it is the model $P(w_j|z_k)$ that describes the relationship between features (visual words) and objects (aspects). Using EM algorithm, it is possible to infer $P(w_j|z_k)$, as we will see later.

Each aspect specifies a mixture of layouts, $\{s_1, ..., s_S\}$, via the factor $P(s_l|z_k)$. Intuitively, a layout encodes the rough locations of objects. For example, s_i can represent that the object is located at the upper-left corner of an image. Using EM algorithm, it is possible to infer $P(s_l|z_k)$, as we will see later.

The position information of visual words is represented by a location dictionary $\{x_1, ..., x_P\}$. The location dictionary 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 $x_p \in \{x_1, ..., x_P\}$.

The location of a visual word is generated by $P(x_p|s_l)$. Since $P(x_p|s_l)$ is combined with priors $P(s_l|z_k)$, we can simply assume $P(x_p|s_l)$ to be a Gaussian distribution for $s \in \{s_1, ..., s_{S-1}\}$. This simple distribution does not restrict the system to model single objects in the scene because the image is composed of a mixture of layouts, each associated with a single-mode distribution, hence being able to model multiple objects as well. To handle the case where no foreground object is present in the scene, we define a layout s_S that corresponds to a uniform distribution; i.e., $P(x_p|s_S) = \text{const}$.

An image d_i is a bag of words and positions. Define the image-word-position co-occurrence table $n(d, w, x)$, with $n(d_i, w_j, x_p)$ denoting the number of occurrences of word w_j

at position x_p in image d_i. This is the table of observations; the other two variables s and z in Fig. (2) are latent.

2.3. The Learning Algorithm

To estimate the factors that maximize the likelihood of the joint probability, we employ the standard Expectation Maximization (EM) algorithm [8]. 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. The goal is to maximize the log-likelihood,

$$\mathcal{L} = \sum_i \sum_j \sum_p n(d_i, w_j, x_p) \log p(d_i, w_j, x_p) \quad (2)$$

Derivations of the E- and M-step are omitted with only the results stated below:

E-step:

$$P(s_l, z_k|d_i, w_j, x_p) \propto P(z_k|d_i)P(w_j|z_k)P(s_l|z_k)P(x_p|s_l) \quad (3)$$

M-step:

$$P(w_j|z_k) \propto \sum_i \sum_p n_{ijp} \sum_l P(s_l, z_k|d_i, w_j, x_p) \quad (4)$$

$$P(z_k|d_i) \propto \sum_j \sum_p n_{ijp} \sum_l P(s_l, z_k|d_i, w_j, x_p) \quad (5)$$

$$P(s_l|z_k) \propto \sum_i \sum_j \sum_p n_{ijp} P(s_l, z_k|d_i, w_j, x_p) \quad (6)$$

$$P(d_i) \propto \sum_j \sum_p n_{ijp} \quad (7)$$

where $n_{ijp} \equiv n(d_i, w_j, x_p)$. Note that these equations need normalization to make them probability distributions. In summary, given $n(d_i, w_j, x_p)$, maximum likelihood fitting by the EM algorithm yields $P(w_j|z_k)$, $P(z_k|d_i)$, $P(s_l|z_k)$, and $P(d_i)$.

2.4. The Inference Algorithm

Given a test image d_{query}, the factors in Eq.(3)$\sim$(7) are computed using the "fold-in" technique described in [2]; 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.

3. LAYOUT EXTRACTION EXPERIMENT

The goal of this experiment is to specify the composition of each image, i.e., where the unknown objects are located. In our experiments, we use 100 face images and 100 non-face images and set the number of latent aspects to two. We use the same set of images as [7], where the face images are taken from the Caltech face dataset [9]. As in [7], images are resized to around 200×140 and converted to grayscale. A random subset of 50 face and 50 non-face images are selected as training images. None of the images contain information about whether it contains faces, nor about where the face is located.

The visual words are computed as explained in Section 2.1. Note that the visual words are neither obtained from labelled data, nor are they specifically designed for this face/non-face task, implying their generality.

We obtain the initial Gaussian means and variances by fitting PLSA to the data, after which we obtain weightings indicating how likely a visual word belongs to an 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 set $S = 4$ in our experiments, but the results are not sensitive to this value. We simply use the average of the weighted variance of all images as the variance $\sigma_l = \sigma$.

During inference, by labelling each visual word with its most likely latent aspect, we can obtain a segmentation-like image. This image indicates the location of the foreground object (see Fig. 3).

The most likely aspect of each visual word can be obtained by first deciding whether a foreground object is present or not: To decide the presence/absence of the foreground object in the scene, we compute

$$P(d_i|z_{\text{face}}, s \neq s_S) = \alpha P(s \neq s_S|z_{\text{face}})P(z_{\text{face}}|d_i)P(d_i) \quad (8)$$

where α is the normalization constant. The larger the value of $P(d_i|z_{\text{face}}, s \neq s_S)$ is, the more likely that image contains a foreground object. Then, if a foreground object is discovered (based on a threshold of equal error rate), we determine the optimal label of each visual word by

$$z^* = \arg\max_z P(z|d_i, w_j, x_p, s \neq s_S). \quad (9)$$

Otherwise, we determine the optimal labels by

$$z^* = \arg\max_z P(z|d_i, w_j, x_p, s = s_S). \quad (10)$$

Once we obtain the optimal labels of the visual words, we create a mask by placing a 1 on each visual word with label $z = z_1$ and a -1 on each visual word with label $z = z_2$. We convolve this mask with a Gaussian to superimpose the labelling within a neighborhood. We threshold this mask to a binary mask and apply the mask on the original image. The results are shown in Fig. 3. Notice that both methods are able to extract the aspect which corresponds to the notion of human face without supervision. Comparing PLSA (column 2) to our method (column 3), it can be seen that foreground objects are more correctly located by our method. Also, PLSA falsely labels many background regions as foreground. PLSA has FAR $= 26.3\%$ and FRR $= 29.2\%$, while our method has a much lower FAR $= 12.4\%$ and FRR $= 15.7\%$, where FAR and FRR are computed based on pixel-level correctness.

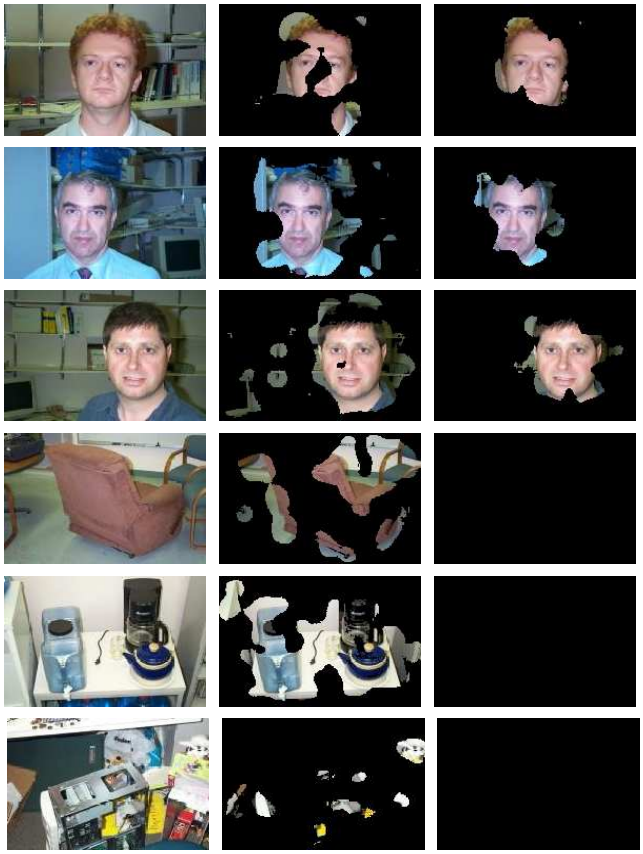

Fig. 3. Second column: result of PLSA. Last column: proposed method.

4. DISCUSSION

Traditional image segmentation can be roughly classified into methods based on clustering and methods based on model fitting [10]. One exemplar of the model fitting approach is Blobworld [1]. In Blobworld, in order to segment an image automatically, the joint distribution of color, texture, and position features is modelled with a mixture of Gaussians. The Expectation-Maximization (EM) algorithm [8] estimates the parameters of this model; the resulting pixel-cluster memberships then provide a segmentation of the image. Traditional image segmentation is performed ***per image*** (with the exception of shape based segmentation methods [11], which are however supervised methods) and the pixel-cluster memberships are not shared across different images. One consequence is that the system cannot gain knowledge from multiple images. In contrast, the shared information is represented in the form of $P(w_j|z_k)$, as explained in Section 2.4.

In the proposed graphical model, the extra dependency between topics and word positions introduces extra variables. Still, the EM algorithm converges stably, and experimental results verify the advantage of the model over Probabilistic Latent Semantic Analysis.

In summary, the significance of our method lies in the capability to perform the whole task in an unsupervised fashion. Although only demonstrated on face images, we expect the method to be also applicable to other objects as in [5].

The current framework uses one Gaussian to model one cluster, but it does not explicitly consider the geometric relationship between visual words *within* each cluster. Extending the framework upon this point is of future interest.

5. ACKNOWLEDGEMENT

This work is supported by the Taiwan Merit Scholarship TMS-094-1-A-049 and by the ARDA VACE program.

6. REFERENCES

[1] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1026–1038, 2002.

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

[3] 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.

[4] 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.

[5] D. Liu and T. Chen, "Semantic-shift for unsupervised object detection," in *IEEE Computer Vision and Pattern Recognition Workshop on Beyond Patches*, 2006.

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

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

[8] 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.

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

[10] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*, Prentice Hall, 2002.

[11] K.M. Pohl, J. Fisher, R. Kikinis, W.E.L. Grimson, and W.M. Wells, "Shape based segmentation of anatomical structures in magnetic resonance images," in *ICCV workshop on Computer Vision for Biomedical Image Applications*, 2005.