SIDOP 2nd Workshop on Signal and Document Processing
March 23-24, 2012 - Hammamet - Tunisia

Topic models for automatic image annotation

Marouane Ben Haj Ayech
LSTS laboratory
ENIT
Tunis, Tunisia
marouane.ayech@yahoo.fr

Faouizi Benzarti
LSTS laboratory
ENIT
Tunis, Tunisia
benzartif@yahoo.fr

Amiri Hamid
LSTS laboratory
ENIT
Tunis, Tunisia
hamidlamiri@yahoo.com

Abstract— **Modern image retrieval systems, which allow users to use textual queries and perform content-based image retrieval (CBIR), depend greatly on Automatic Image Annotation. Many models namely unsupervised topic models have successfully been applied in text analysis and are showing encouraging results in automatic image annotation. In this work, we first describe the basic topic models: the latent semantic analysis (LSA), the probabilistic latent semantic analysis (PLSA) and the latent Dirichlet allocation (LDA). Since these models assume that documents are represented as "bag of words" in text analysis, we then describe BOV-based image representation, an analogous representation adapted to image annotation. Based on SIFT technique followed by vectorial quantization, this representation allow image to be as "bag of visterms" (BOV). Finally, we describe some advanced topic models: GM-PLSA, GM-LDA and CORR-LDA, which are used in image annotation.**

Keywords- automatic image annotation; topic models ; CBIR ; SIFT ; BOV.

I. INTRODUCTION

Automatic image annotation, which means the association of words to whole images [1], became a crucial part of information retrieval systems especially content-based image retrieval ones. Indeed, users prefer using textual request when searching images. However, retrieving within CBIR involves using low-level visual features of images. So, we fall in the semantic gap problem. To resolve this problem, almost all researchers tried to implement efficient techniques of image annotation. These approaches must annotate images automatically to treat large image databases.

Many approaches have been proposed for semantic image annotation and retrieval and are roughly classified into two categories: supervised versus unsupervised approaches [9].

The first class treats the annotation problem as a supervised classification and the words as independent classes. An important principle of these methods is to perform similarity measure at the visual low-level and annotate unseen images by propagating the corresponding words. The most important works found in the literature are Chang et al., 2003[5] and Carneiro et al., 2007[6].

The second class treats images and texts as equivalent data. Thus, they apply an unsupervised learning over data in order to discover the correlation between visual features and textual words. So, the annotation is posed as statistical inference which treats images as a bag of words and features generated both by latent or hidden variables. Various models have adopted this idea. Mori et al propose a model that use co-occurrence between words and features to predict words for annotating unseen images. Duygulu et al propose a translation model between two languages one for blobs and another for words that translates blobs into words, i.e., it attaches words to new image regions. Lavrenko et al propose the continuous-space relevance model (CRM) in which word probabilities are estimated using multinomial distribution and the blob feature using a non-parametric kernel density estimate [9]. Some other models that associate latent aspects or topics with images are called topic models such as LSA, PLSA and LDA and are successfully used in text analysis. So, in this work, we describe them and their extensions which are designed for image annotation.

II. RELATED WORK

Topic models require that data representation is based on Bag-of-Words model, which implies that spatial relationships between words are ignored. Thus, the common way followed is to represent data as an observation matrix noted *A* that we will explain in detail later.

The idea behind these models is to add levels of latent variables to model aspects or topics. Since they were designed to be applied in text analysis, we prefer keep the terminology used in text analysis and after that we present analogy with image annotation in section 3. In this section, we are interested in the following simple models: LSA, PLSA and LDA.

LSA is Linear Algebra-based model and exploits data, i.e. matrix A, from algebraic perspective, while PLSA and LDA

are statistical models and belong to the class of probabilistic generative models.

A. Bag of Words (BOW)model

This model is called also "Orderless Document Representation". Indeed, it assumes that order of words has no significance; i.e., the term "home made" has the same probability as "made home".

BOW model simplifies data representation and when applied to a corpus of text, gives a simple data representation.

Suppose we have a corpus C that is a collection of documents $C = \{d_1, \dots, d_N\}$. Each document d_i consists of a set of words. C is represented by a term-by-document matrix $A \in \mathbb{R}^{N\times M} = (a_{i,j})_{\mathrm{i=1..N,j=1..M}} = \mathrm{n(d_i, w_j)}$, where N is the number of documents, M is the vocabulary size and $\mathrm{n(d_i, w_j)}$ is the number of occurrences of w_j in document d_i. The vocabulary V is a set of all possible words in the corpus $V = \{w_1, \dots, w_N\}$.

B. Latent Semantic Analysis (LSA)

LSA is a Linear Algebra-based model which consists in decomposing the term-by-document matrix using Singular value decomposition (SVD) [10].

$$A \cong USV^T \quad (1)$$

where $A \in \mathbb{R}^{N\times M}, U \in \mathbb{R}^{N\times K}, S \in \mathbb{R}^{K\times K}$ and $A \in V^{K\times M}$.

LSA consists in projection of the original space onto reduced dimensionality space, which allows capture of hidden similarity between terms. Unfortunately, LSA lacks a probabilistic interpretation [1].

C. Probabilistic Latent Semantic Analysis (PLSA)

In this model, a document is not represented as a bag of words but is modeled as a mixture of aspects or topics. Each topic is represented as a multinomial words distribution [11].

This model is based on a conditional independence assumption: Each observed word w_j is conditionally independent of the document d_i it belongs to given a latent variable z_k.

$$P(w_j, d_i) = P(d_i)P(w_i|d_j) \quad (2)$$

$$P(w_i|d_j) = \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \quad (3)$$

Since the number of latent variables is smaller than the number of words or documents, z_k behave as a bottleneck in predicting words.

Model fitting is performed using EM algorithm, which alternates two steps to assure maximum likelihood estimation:

E-step: The conditional distribution $P(z_k|d_i, w_j)$ is computed from the previous estimate of parameters:

$$P(z_k|d_i, w_j) = \frac{P(z_k|d_i)P(w_j|z_k)}{\sum_{l=1}^{K} P(z_l|d_i)P(w_j|z_l)} \quad (4)$$

M-step: The parameters $P(z_k|d_i)$ and $P(w_j|z_k)$ are updated with the new expected values $P(z_k|d_i, w_j)$:

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i,w_j)P(z_k|d_i,w_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i,w_m)P(z_k|d_i,w_m)} \quad (5)$$

$$P(z_k|d_i) = \frac{\sum_{i=1}^{N} n(d_i,w_j)P(z_k|d_i,w_j)}{\sum_{j=1}^{M} n(d_i,w_j)} \quad (6)$$

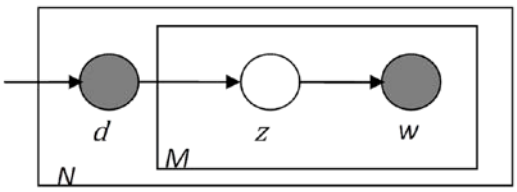

Figure 1. The graphical model of PLSA. Nodes inside a given box indicate that they are replicated the number of times indicated in the bottom left corner. Filled circles indicate observed random variables; unfilled are unobserved.

The maximum likelihood estimation is released by maximizing the objective function :

$$L = \prod_{i=1}^{N}\prod_{j=1}^{M} P(w_i|d_j)^{n(w_i,d_j)} \quad (7)$$

where $P(w_i|d_j)$ is given by (3).

D. Latent Direchlet Allocation (LDA)

In contrast to PLSA, LDA treats the multinomial weights over topics as latent random variables. Those weights are sampled from a Dirichlet distribution, which is the conjugate prior of the multinomial distribution [8]. Since LDA is a combination of two conjugate distributions, it normally has fewer parameters than PLSA model [4] and by consequent reduce overfitting.

LDA assumes the following generative process:

1. Choose $N \sim$ Poisson(ξ)
2. Choose $\theta \sim$ Dir(α)

$$P(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}\theta_i^{\alpha_i-1}$$

3. For each of the N words w_n

   (a) Choose a topic $z_n \sim$ Multinomial(θ)

   $$P(z_n|\theta) = \theta_i = \prod_{l=1}^{k}\theta_i^{w^l}$$

   where $w^l=1$ if $l=i$ and $w^l=0$ if $l\neq i$

   (b) Choose a word w_n from $p(w_n \mid z_n;\beta)$, a multinomial probability conditioned on the topic z_n.

   $$P(w_n|z_n,\beta) = P(w^j = 1|z^i = 1) = \beta_{ij}$$

where N is the number of words contained in a document, $\boldsymbol{w} = (w_1, \dots, w_n, .., w_N)$ is the document representation.

$w_n = (w_n{}^1, \dots, w_n{}^j, \dots, w_n{}^V)$ is the representation of the n[th] word in $\boldsymbol{w}$, where $w_n{}^j = 1$ and $w_n{}^l = 0$ if $l \neq j$

$z_n = (z_n{}^1, \dots, z_n{}^i, \dots, z_n{}^K)$ is the representation of the n[th] word in $\boldsymbol{z}$, where $z_n{}^i = 1$ and $w_n{}^l = 0$ if $l \neq i$

We note that:

- The three-level hierarchical structure of LDA allows that many concepts may be associated to one document.
- The words are generated by concepts

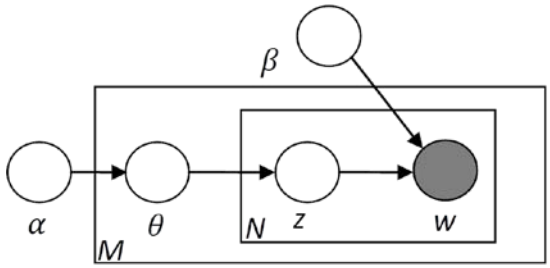

Figure 2. The graphical model of LDA

To estimate the parameters of LDA, we must apply inference and estimation procedures to the model. Thus, we need to compute the posteriori distribution in inference step:

$$P(\theta, z|\boldsymbol{w}, \alpha, \beta) = \frac{P(\theta, z, \boldsymbol{w}|\alpha, \beta)}{P(\boldsymbol{w}|\alpha, \beta)} \quad (8)$$

where

$$P(\theta, z, \boldsymbol{w}|\alpha, \beta) = P(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)\, p(w_n|z_n, \beta)$$

$$P(\boldsymbol{w}|\alpha, \beta) = \int P(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta)\, p(w_n|z_n, \beta) \right) d\theta$$

The posterior distribution is intractable and the solution is to use variational estimation methods [8].

III. BOV-BASED IMAGE REPRESENTATION

Since aspect models are firstly used in text analysis, we have used text terminology (document, word, ..). We present now the analogy of this terminology with image annotation domain.

Thus, images represent documents and words correspond to visterms. In fact, an image is divided into regions (visterms).

Now, we describe briefly the procedure used for image representation which allows representing an image as set of visterms:

First, we apply an interest-point detector on each image to extract characteristic points. There exist in the literature many techniques such as DoG (the difference of Gaussians) point detector. The chosen technique must detect invariant points to some geometric and photometric transformations. Then, we apply SIFT (Scale Invariant Feature Transform) to obtain local descriptors from each image. These descriptors are computed on the regions around each interest point identified by the detector. Since the descriptors are obtained and in order to get a fixed image representation, we quantize all descriptors into a discrete set of visterms using by example k-means. Each cluster obtained represents a visterm in the image.

A. Scale Invariant Feature Transform (SIFT)

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

SIFT is a method for extracting distinctive and invariant features from images that can be used to perform reliable matching between different views of an object or scene [3]. The stages of computation used to generate the set of image features are:

- Scale-space extrema detection: the first stage uses an interest-point detector; difference of Gaussians (DOG) which is a function to identify potential interest points that are invariant to scale and orientation.
- Keypoint localization: at each candidate location, a detailed model is fit to determine location and scale. Keypoints are selected based on measures of their stability.
- Orientation assignement: One or more orientations are assigned to each keypoint location based on local image gradient directions.
- Keypoint descriptor: The local image gradients are measured at the selected in the region around each keypoint. These are transformed into a representation that allows for significant levels of local shape distorsion and change in illumination.

B. Vectorial Quantization

Once SIFT has been applied to the collection of images, we obtain a set of feature vectors corresponding to images regions. We use k-means, an unsupervised quantization technique, and we apply it over the whole feature space. The k-means algorithm yields to a number of centroids (their number must be fixed before applying k-means). Each centroid is a vector whose length equals to feature space dimension and is representative to a subset from feature space. These centroids will be the visterms required by BOV model.

C. Bag-of-visterms

An image is modeled using the Bag-of-visterms model, which is a simple model that represents a document as an orderless set of terms. In the case of images, an image is

therefore represented as a orderless sequence of visual terms, called visterms.

Given a collection of images, the first task to perform is to identify a set of all visterms used at least once in at least one image. This set is called the vocabulary. Although the image is a set, we fix an arbitrary ordering for it so we can refer to visterm1 through visterm M where M is the size of vocabulary. Once vocabulary has been fixed, each image is represented as a vector with integer entries of length M. If this vector is d then its j^{th} component d_j is the number of appearances of visterm j in the image. The length of image is $n = \sum_{i=1}^{M} d_j$.

As seen above, the collection of images is represented as a N-by-M matrix, where each row describes an image and each column corresponds to a visterm.

IV. TOPIC MODELS FOR IMAGE ANNOTATION

In this section, we describe three advanced topic models: Gaussian-Multinomial PLSA, Gaussian-Multinomial LDA and Correspondence LDA. These three models are based on basic topic models seen above: PLSA and LDA. Given that these basic models are suitable only for modeling one-type of data, advanced topic models are designed to fit multi-type data.

A. Gaussian-Multinomial PLSA (GM-PLSA)

GM-PLSA is a combination of two PLSA models: a standard PLSA to model textual words and a continuous PLSA to model visual features. These two models share a common distribution over latent variable *z* noted P(*z*|*d*).

The whole model, which is represented in figure 3, assumes the following generative process:

1. Select a document d_i with probability $P(d_i)$
2. Choose a latent aspect z_k with probability $P(z_k|d_i)$ from a multinomial distribution conditioned on dicument d_i
3. For each of the words, sample w_m from a multinomial distribution $Mult(\theta_k)$ conditioned on the latent aspect z_k
4. For each of the feature vectors, sample r_n from a multivariate Gaussian distribution $\mathcal{N}(x|\mu_k, \sigma_k)$ conditioned on the latent aspect z_k

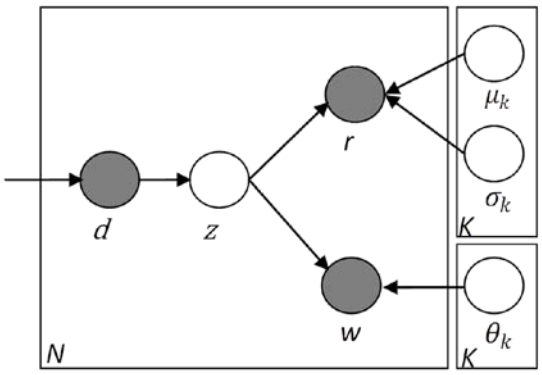

Figure 3. The graphical model of GM-PLSA

B. Gaussian-Multinomial LDA (GM-LDA)

GM-LDA is a combination of two LDA models: a standard LDA to model textual words and a continuous LDA to model visual features. This model, represented in figure 4, shows that words w_m and regions r_n of an image can come from different topics which means that the whole document can contain multiple topics. Furthermore, we can view θ as high-level representation of the whole document (image features + words).

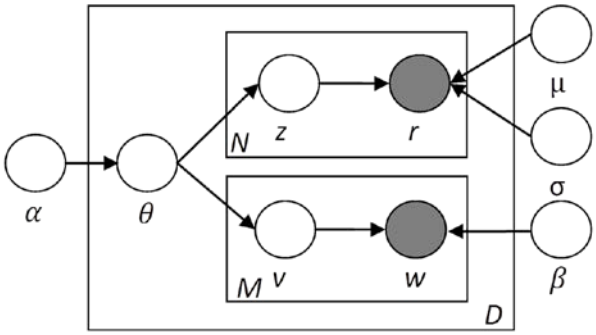

Figure 4. The graphical model of GM-LDA

The joint probability distribution of the set of image features *r*, the set of associated words *w* and latent variables θ, *z* and *v* is given as follows:

$$p(\boldsymbol{r}, \boldsymbol{w}, \theta, \boldsymbol{z}, \boldsymbol{v}) = p(\theta|\alpha)\left(\prod_{n=1}^{N} p(z_n|\theta)p(r_n|z_n, \mu, \sigma)\right)\left(\prod_{m=1}^{M} p(v_m|\theta)p(w_n|v_m, \beta)\right)$$

GM-LDA model assumes the following generative process:

1. Sample a Dirichlet random variable θ
2. For each of the N image regions:
   a. Sample $z_n \sim Mult(\theta)$
   b. Sample a region description r_n conditional on z_n
3. For each of the M words:
   a. Sample $v_m \sim Mult(\theta)$
   b. Sample a word w_m conditional on v_m

C. Correspondance- LDA (CORR-LDA)

In CORR-LDA model, image features are firstly generated and subsequently words are generated. Indeed, *N* region features are generated. Then, for each of the *M* words, one of the regions is selected from the image and a corresponding word is drawn conditioned on the topic that generates the selected region (Figure 5).

The joint probability distribution of the set of image features *r*, the set of associated words *w* and latent variables θ, *z* and *y* is given as follows:

$$p(\boldsymbol{r}, \boldsymbol{w}, \theta, \boldsymbol{z}, \boldsymbol{y})$$
$$= p(\theta|\alpha)\left(\prod_{n=1}^{N} p(z_n|\theta)p(r_n|z_n,\mu,\sigma)\right)\left(\prod_{m=1}^{M} p(y_m|N)p(w_n|y_m,\boldsymbol{z},\beta)\right)$$

CORR-LDA model assumes the following generative process:

1. Sample $\theta \sim Dir(\theta|\alpha)$
2. For each of the N image regions:
   a. Sample $z_n \sim Mult(\theta)$
   b. Sample a region description r_n from a multivariate Gaussian distribution conditional on z_n
      $r_n \sim P(r|z_n,\mu,\sigma)$
3. For each of the M words:
   a. Sample $y_m \sim Unif(1,\dots,N)$
   b. Sample a region description w_m from a multinomial distribution conditional on y_m and $\boldsymbol{z}$

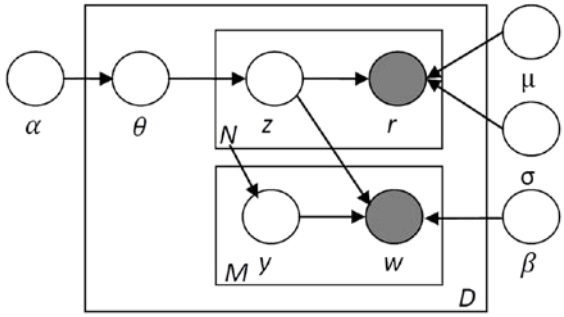

Figure 5. The graphical model of CORR-LDA

In this paper, we have described topic models. The basic models, PLSA and LDA are designed for one-type data and advanced models are designed for modeling multi-type data, especially for image modeling and annotation.

REFERENCES

[1] F. Monay and D. Gatica-Perez, "On Image Auto-Annotation with Latent Space Models", ACM Multimedia (2003) , p. 275--278

[2] K. Barnard, P. Duygulu, N. Freitas, D. Forsyth, D. Blei, "Matching words and pictures," Journal of Machine Learning Research, 3:1107-1135, 2003.

[3] D.G.Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", International Journal of Computer, 2004.

[4] D.M. Blei, M.I. Jordan, "Modeling annotated Data", In Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 127–134. ACM Press, 2003.

[5] E. Chang, K. Goh, G. Sychay and G. Wu, CBSA: Content-based soft annotation for multimodal image retrieval using Bayes point machines. IEEE Trans. Circ. Systems Video Technol., **13** 1 (2003), pp. 26–38.

[6] G. Carneiro, A.B. Chan, P.J. Moreno and N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans. Pattern Anal. Machine Intell., 29 3 (2007), pp. 394–410.

[7] S. Tollari, "Image indexing and retrieval by combining textual and visual informations", thesis in Université du Sud Toulon-Var, 2006.

[8] D. M.Blei,A. Y.Ng,M. I.Jordan, "Latent Dirichlet Allocation". Journal of Machine Learning Research, 3, 993-1022,2003.

[9] Z. Li, Z. Shi, X. Liu, Z. Shi, "Modeling continuous visual features for semantic image annotation and retrieval", 2011

[10] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "Indexing by latent semantic analysis", Journal of the American Society of Information Science, 41(6):391-407, 1990

[11] T. Hoffman, "Probabilistic latent semantic indexing", SIGIR Conference, 1999

BIBLIOGRAPHY

Ben Haj Ayech Marouane received an engineering degree from ENIT in 2007and the M.Sc. degree from ENIT in 2009. He is now a phd student at ENIT, His research interest includes information retrieval, machine learning and computer vision.

Hamid Amiri received the Diploma of Electrotechnics, Information Technique in 1978 and the PhD degree in 1983 at the TU Braunschweig, Germany. He obtained the Doctorates Sciences in 1993. He was a Professor at the National School of Engineer of Tunis (ENIT), Tunisia, from 1987 to 2001. From 2001 to 2009 he was at the Riyadh College of Telecom and Information. Currently, he is again at ENIT. His research is focused on

- Image Processing.
- Speech Processing.
- Document Processing.
- Natural language processing