Automatic Image Annotation through Multi-Topic Text Categorization

Sheng Gao[1]*, De-Hong Wang*[1,2] *and Chin-Hui Lee*[3]

[1]Institute for Infocomm Research, Singapore 119613
[2]National University of Singapore, Singapore 117543
[3]School of Electrical & Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
{gaosheng, dehong}@i2r.a-star.edu.sg chl@ece.gatech.edu

ABSTRACT

We propose a new framework for automatic image annotation through multi-topic text categorization. Given a test image, it is first converted into a text document using a visual codebook learnt from a collection of training images. Latent semantic analysis is then performed on the tokenized document to extract a feature vector based on a visual lexicon with its vocabulary items defined as either a codeword or a co-occurrence of multiple codewords. The high-dimension feature vector is finally compared with a set of topic models, one for each concept to be annotated, to decide on the top concepts related to the test image. These topic classifiers are discriminatively trained from images with multiple associations, including spatial, syntactic, or semantic relationship, between images and concepts. The proposed approach was evaluated on a Corel dataset with 374 keywords, and the TRECVID 2003 dataset with ten selected concepts. When compared with state-of-the-art algorithms for automatic image annotation on the Corel test set our system obtained the best results, although we only use a simple linear classification model based on just texture and color features.

1. INTRODUCTION

A large volume of digital image/video documents is becoming available on the web and in digital libraries. This gives rise to research opportunities related to organizing, indexing and searching of multimedia documents. However, the inherent complexity in representing and recognizing images make it more difficult to deal with than text documents. Unlike spoken and text documents, a semantic description of an image or a video segment is usually unavailable. One way to circumvent these limitations is to associate semantic concept annotations with image contents. This can usually be accomplished through manual annotation, a non-trivial, if not impossible, task. Recently, some techniques for automatic image annotation (AIA) have been developed to address some the issues [1-3, 5, 7, 10-14].

Most of the recent studies learn the statistical models, which characterize the joint statistical distribution between the observed visual features and annotated keywords, for annotating images. Prominent examples are cross-media relevance model (CMRM) [3], translation model (TM) [5, 7], multiple Bernoulli relevance model (MBRM) [11], continuous-space relevance model (CRM) [12], maximum entropy (ME) [2], Markov random field (MRF) [6], and conditional random fields (CRF) [10, 14]. An image is divided into a set of sites or elements (i.e. grids or regions) and the visual features extracted from the grids are used for representing images. For CRM and MBRM models, the Gaussian densities are applied to modeling distributions of visual features, while others apply *k*-means clustering to tokenize the image elements. The spatial dependency among image elements is often not modeled, except for MRF. But MRF only models the nearest neighborhood dependency because its parameter size increases exponentially while modeling long-span dependencies. CRF is able to characterize the long-span dependency as studied in [14] for image annotation. But MRF and CRF incur high computation cost and in some cases infeasible inferences.

Learning from the state-of-the-art algorithms in automatic speech recognition and text data mining, it is desirable to characterize image content with a set of visual symbols and represent it as a text document. So many techniques, commonly used in information retrieval for statistical modeling and classifier learning, are directly applicable to AIA. The obvious benefit is that modeling of syntactic and semantic relationship among the visual symbols can be explored. However, this symbolic representation, naturally existing in text documents, is not readily available for images. A key challenge is the tokenization of image features into a set of visual alphabets, so that language modeling of these alphabets is utilized without an explicit association of any linguistic description of these symbols.

In this paper, we propose a unified framework for AIA. An image is first tokenized using a visual codebook, and its content is represented with a high-dimensional feature vector as discussed in Section 2. Then AIA is then abstracted in Section 3 as a multi-category (MC) text categorization (TC) problem. As shown in [9] that MC maximal figure-of-merit (MFoM) learning is a powerful tool for designing discriminative classifiers, it is adopted here to train concept models for image annotation. Evaluation is carried out on the Corel dataset with 374 keywords, and the TRECVID 2003 dataset with a selected set of ten concepts. The AIA results are compared with a few state-of-the-art algorithms in Section 4. Finally we summarize our findings in Section 5.

2. TEXT REPRESENTATION OF IMAGES

Image representation has been a topic of intensive study in the image processing and computer vision communities. For semantic content description, it is desired to associate and annotate a given image with a set of multiple concepts describing the objects and their spatial relations in an image. But such detectors are generally unavailable. Next we will introduce an approach to describing the content of an image as a text document.

2.1 Tokenization of Image Content

A text document is described by a sequence of words defined in a lexicon. For text categorization, the popular way is to view the document as a “bag-of-words”, i.e. the order of words is ignored. Then a high-dimensional vector is extracted to represent the document. This vector will encapsulate the statistics (e.g. co-occurrence of semantic and syntactic relations) of occurred terms in the document. However, the set of visual symbols needed to accomplish this representation is usually unavailable.

1-4244-0469-X/06/$20.00 ©2006 IEEE

To make use of such a text representation, the first step is to choose a set of image alphabets (or visual terms) and build a visual lexicon based on these terms and their co-occurrences. An ideal visual term in the lexicon should carry semantic meanings. Intuitively, they may be objects in the image describing some semantic concepts. Unfortunately, developing generic object or semantic concept detectors is still a challenging problem, although some special detectors (e.g. face detector) can be accessed in some specific scenarios. Here we apply a feasible method that uses unsupervised clustering algorithms to automatically learn a visual lexicon. In [2-3, 5, 7], *k*-means clustering is used to get a set of tokens, each describing a cluster of sub-blocks. Although these tokens represent vague meanings, especially for the grid-based clustering, its success for AIA has been demonstrated. Here we further extend this method to: (1) learning a collection of visual terms, each of which may be a single token or a combination of tokens (e.g. pair-wise tokens); (2) learning an ensemble of visual lexicons (e.g. color lexicon, texture lexicon) from the different sets of low-level visual features, each describing a partial content of an image; and (3) learning co-occurrence statistics of visual terms to enhance the representation.

Different from the conventional object-based features, a regular segmentation of an image into I by J blocks is first preformed. Then the low-level, raw image features are extracted and grouped into a feature vector for each block. For example, an image is first divided into a collection of blocks of 16x16 pixels each. Low-level image features, such as color histograms, textures and DCT coefficients, are obtained from the block. These raw features are grouped into a vector, X_{ij}, for the block located at the *i*-th row and *j*-th column. All vectors from training images are quantized to form a codebook of N tokens, with each block tokenized as an index in the codebook after quantization.

2.2 Learning Ensemble Visual Lexicons

The visual terms in a lexicon consist of not only the tokens but also any pattern inferred from the token relations (e.g. location, spatial, *n*-gram, etc.). If an object detector is available, it can also be used to tokenize the image into more meaningful semantic and syntactic relations. For example, if “sky” and “sea” are detected, one additional visual term, *“sky” is above “sea”*, can be extracted to describe this relation. Since each visual lexicon only describes a partial content, an ensemble visual lexicon is desirable to enhance the power of representation. It is known that simple concatenation of them does not work well. The ensemble lexicons will address the issue because each lexicon is independently learned from the distinctive feature. There are many ways to learn the visual lexicon. Here one possible way used in our experiment is described as follows.

We extend the symbolic representation to incorporate the contextual or spatial dependency into an image pattern. Here the pattern means a symbol sequence such as *n*-gram or a combination of the symbols according to some syntactic rules. A visual lexicon is constructed using all detected patterns. Figure 1 shows a possible way. For block X_{22}, its direction-specific bigram patterns, such as $X_{22}X_{21}$, $X_{22}X_{23}$, $X_{22}X_{11}$, $X_{22}X_{33}$, are obtained from its neighboring blocks. Here 8 directions are shown. The extracted bigrams are treated as distinctive patterns. Sometimes these patterns are further clustered to reduce the size of the visual lexicon. The patterns function similarly to the terms or words in text documents. If multiple visual features are available, multiple visual lexicons can be built. The ensemble lexicons are utilized to represent image content. So an image is viewed as a text document with the combinational representation by the terms in the ensemble lexicons. Besides *n*-gram statistics, other patterns can be explored, e.g. cross visual lexicon patterns as in [13], and syntactically related patterns.

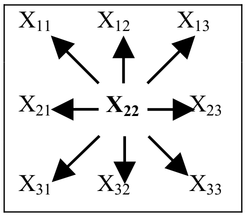


Figure 1 An example to show bigram visual terms

2.3 Image-Text Representation

After an image is tokenized and the occurrence statistics of visual lexicons are tabulated, a feature vector is extracted for content representation using techniques developed in IR [8]. For example, given a color lexicon, $A=\{A_1,A_2, ..., A_M\}$, with M visual color terms, the color content of an image is represented by a vector, $V=(v_1,v_2,\cdots,v_M)$, each component being the statistics of the visual term occurred in the image. For instance, v_i is the statistic for the term A_i. The vector dimension is usually high. For a visual lexicon, only having unigram and bigram patterns, its dimension should be $M+M*M$. By using a 64-token codebook, its dimension reaches 4160. To reduce the dimension, we can remove the frequently occurred or very rare patterns, or apply the feature selection and reduction techniques. Here we introduce the latent semantic indexing (LSI) technique used in all of our experiments.

Given a training set, T, with the size K, the vector, V^j, for the *j-th* image, is calculated as in [4], whose *i-th* component is defined as,

$$v_i^j = (1-\varepsilon_i)\cdot c_i^j \big/ n^j \qquad (1)$$

where c_i^j is the number of times of A_i occurred in the *j-th* image, n^j is the total number of visual terms appeared in the *j-th* image, and ε_i is a normalized entropy of A_i defined as,

$$\varepsilon_i = -\frac{1}{\log K}\sum_{j=1}^{K}\frac{c_i^j}{t_i}\log\frac{c_i^j}{t_i} \qquad (2)$$

where $t_i=\sum_j c_i^j$ denotes the total occurrence count of A_i. Then a term-document matrix, D_{MxK}, is constructed, whose *j-th* column is V^j. Dimension reduction is done by selecting only the top-N eigenvectors after singular value decomposition (SVD) [4].

3. IMAGE ANNOTATION

Now we describe building concept models using the MC MFoM approach [9]. MC MFoM learns multi-category classifier by optimizing a metric-oriented objective function. It is more robust and works better than the popular SVM classifiers, especially for learning in the case with sparse training [9], which frequently occurs in AIA.

3.1 Automatic Image Annotation

In AIA, a training image set, $T=\{(X,Y)\big|X\in R^D, Y\subset C\}$, is given, where *(X, Y)* is a training sample. X is a D-dimensional feature extracted as discussed in Sections 2. Y is the manually assigned annotation with multiple keywords or concepts. The predefined keyword set is denoted as $C=\{C_j, 1\leq j\leq N\}$, with N the total number of keywords and C_j the j-th keyword. Clearly Y is a subset of C. We will learn a discriminant function with the parameter set Λ_j, $g_j(X;\Lambda_j)$, for the *j-th* keyword from *T*.

In the evaluation stage, multiple relevant keywords are assigned to an image *X*, according to the following multiple-label decision rule,

$$\begin{cases}\text{Accept} \quad \mathrm{X}\in C_j \ \text{ if } g_j(X;\Lambda_j)-g_j^-(X;\Lambda^-)>0 \\ \text{Reject} \quad \mathrm{X}\in C_j, \mathit{Otherwise}\end{cases} \quad 1\leq j\leq N, \quad (3)$$

where $g_j^-(X;\Lambda^-)$ is named as *class anti-discriminant function* for the *j-th* keyword, defined as,

$$g_j^-(X;\Lambda^-)=\log\left[\frac{1}{|C_j^-|}\sum_{i\in C_j^-}\exp\left(g_i(X;\Lambda_i)\right)^{\eta}\right]^{1/\eta}, \quad (4)$$

where C_j^- is a subset containing the most competitive keyword models against C_j, $|C_j^-|$ is its cardinality, Λ^- is the parameter set for all competitive keyword models, and η is a positive constant. Eq. (4) measures the score as a geometric average of scores among all competing categories. It functions as a negative model for the *j*-th keyword.

If the size of *C* is large, the cost of verifying all possible decisions in Eq. (3) is high. Fortunately, the nature of multi-category learning makes it possible to compare scores from *N* concept models. So it is enough to verify only the *top-M* keyword candidates. This is one benefit from jointly designing multi-category classifiers. It is well known that the scores from separately learnt binary classifiers cannot be directly compared.

3.2 MC MFoM Learning

In MC MFoM learning [9], the parameter set, $\Lambda=\{\Lambda_j, 1\leq j\leq N\}$, is estimated by optimizing a metric-oriented objective function. The continuous and differentiable objective function, embedding the model parameters, is specially designed for approximating the chosen performance metric (e.g. precision, recall or F_1).

To complete the definition of the objective function, a one-dimensional class misclassification function, $d_j(X;\Lambda)$, is introduced to smooth the discrete decision rule in Eq. (3),

$$d_j(X;\Lambda)=-g_j(X;\Lambda_j)+g_j(X;\Lambda_j^-) \quad . \quad (5)$$

where $d_j(X;\Lambda)<0$ when a correct decision is made. Otherwise, $d_j(X;\Lambda)\geq 0$. It functions similar to Eq. (3). Since Eq. (5) is valued from $-\infty$ to $+\infty$, a class loss function, $l_j(X;\Lambda)$, for the keyword C_j, is further defined for normalization,

$$l_j(X;\Lambda)=\frac{1}{1+e^{-\alpha\left(d_j(X;\Lambda)+\beta\right)}} \quad , \quad (6)$$

where α is a positive constant that controls the size of the learning window and the learning rate, and β is a constant measuring the offset of $d_j(X;\Lambda)$ from 0. They are empirically determined. The value of Eq. (6) simulates the error count made by the *j*-th image model for a given sample *X*.

With the above definitions, most commonly used metrics, e.g. precision, recall and F_1, are approximated over training set *T*:

$$FN_j\approx\sum\nolimits_{X\in T} l_j(X;\Lambda)\cdot 1(X\in C_j) \quad , \quad (10)$$

$$FP_j\approx\sum\nolimits_{X\in T}\left(1-l_j(X;\Lambda)\right)\cdot 1(X\notin C_j) \ , \quad (11)$$

$$TP_j\approx\sum\nolimits_{X\in T}\left(1-l_j(X;\Lambda)\right)\cdot 1(X\in C_j), \quad (12)$$

where TP_j is the true positive, FP_j is the false positive, and FN_j is the false negative for the *j*-th keyword. *1(.)* is an indicator function of any logical expression. In the experiment, the micro-averaging F_1 is our preferred metric. Therefore, the objective function is defined as,

$$L(X;\Lambda)=2\sum\nolimits_{i=1}^{N}TP_i\Big/\left[\sum\nolimits_{i=1}^{N}FP_i+\sum\nolimits_{i=1}^{N}FN_i+2\sum\nolimits_{i=1}^{N}TP_i\right], \quad (13)$$

It is solved using a generalized probabilistic descent algorithm [9].

4. EVALUATION AND RESULT ANALYSIS

The proposed framework is evaluated on the Corel CD [7] and TRECVID 2003 datasets. The Corel set has 374 concepts with a total of 5,000 images, 4,500 images for training and 500 for testing. And the TRECVID set has 33,529 key frames from 93 MPEG files. It is annotated with a set of 114 concepts, out of which only 10 concepts (i.e. *Aircraft including Airplane, Airplane_landing and Airplane_takeoff, Animal, Building, Car/Bus/Truck including Car, Bus and Truck, News subject face including Male_News_Subject and Female_News_Subject, Non-studio setting, Outdoors, People, Road and Weather news.*) were selected in this experiment. We randomly select a half, i.e. 15,804 key frames, for training, and the remaining 17,725 for testing.

For the Corel data, each image is uniformly segmented into 96 grids with a grid size of 16x16, while 77 grids with a size of 32x32 were obtained for TRECVID images. Then a 12-dimensional color feature vector (mean and variance of RGB and LAB value), and a 12-dimensional texture feature vector (energy of log Gabor filter[1]) are extracted from each grid and normalized to zero mean with a unit variance. Then *k-means* clustering is used to get 64 symbols for image tokenization.

4.1 Multi-Category Classifier

A linear classifier is trained using MC MFoM by optimizing the objective function in Eq. (13),

$$g_j(X;\Lambda_j)=W_j\cdot X+b_j \quad (14)$$

where W_j and b_j are the parameters for the *j-th* concept model.

4.2 Comparison with the State-of-the-art

The model based transformation (MBT) method for multi-category classification in [13] was used to fuse the color and texture features. First, two types of classifiers were trained independently on the color and texture features. Then their outputs, a total of 374*2 dimensions for Corel, and 10*2 for TRECVID, were used to train the second classifier. A comparison with the state-of-the-art results is shown in Table 1 based on the Corel set. Table 2 gives the results for only color feature (E1), only texture feature (E2), and fusing with the MBT (E3). All shown results were measured by the averages of precision (*mP*), recall (*mR*), and F_1 (*mF*) over all keywords and the number of concepts detected (i.e. concepts with recall > 0). The published results with TM [12], CMRM [12], ME [2], and MBRM [11] were used for comparison. Because only 260 of the 374 keywords occur in the test set, the results for the set of 260 concepts are compared in Table 1.

[1] http://www.csse.uwa.edu.au/~pk/Research/MatlabFns

It is seen that the proposed framework achieves the best. The *mP* and *mR* reach 0.25 and 0.27, respectively, with 133 concepts detected. Compared with the MBRM model, an improvement was seen. The reason may be that MBRM is a generative model using continuous-valued features, while discrimination may deteriorate because of tokenization. When compared with the other three tokenization-based models, significant improvements were seen.

Table 1 Comparison with state-of-the-art on Corel (260 concepts)

	TM	CMRM	ME	MBRM	Proposed
mP	0.06	0.10	0.09	0.24	**0.25**
mR	0.04	0.09	0.12	0.25	**0.27**
# of det	49	66	N.A	122	**133**

Table 2 Performance on Corel (374 concepts) and TRECVID

		mP	mR	mF	# of det
Corel	E1	0.166	0.130	0.146	99
	E2	0.105	0.102	0.103	88
	E3	**0.171**	**0.188**	**0.179**	**133**
TREC VID	E1	0.249	0.188	0.214	10
	E2	0.194	0.157	0.173	9
	E3	0.196	**0.288**	**0.233**	10

4.3 Annotation Examples

Two images each from the Corel and TRECVID sets were illustrated in Figure 2 as annotation examples. For each image, its identity number (*ID*), ground truth (*Truth*), annotation results using color (*E1*), texture (*E2*), and fusion (*E3*) features, are listed respectively. Different from other annotation algorithms that use fixed-length labels to annotate images, our proposed algorithm obtained sets of annotation results of variable sizes according to the confidences of keyword models (see Eq.(3-4)).

5. CONCLUSION AND FUTURE WORK

We propose a new framework for multi-category automatic image annotation. An image is first tokenized using the visual codebook, and a visual lexicon is built from the spatial, syntactic, semantic associations of tokens. Then the co-occurrence statistics of the visual terms are utilized to characterize its content. Due to the high dimension and sparse training of image samples, MC MFoM is adopted to train concept models. We report experimental results on the Corel set and TRECVID 2003 set. The proposed framework achieved the best results on the Corel set with an average of 0.25 in precision, and 0.27 in recall, and detected 133 concepts. In the future, we will extract more meaningful tokens for quantization, for example, clustering based on salient point detectors, and exploring more relations for mining visual patterns.

6. REFERENCES

[1] D. Blei and M.-I. Jordan., “Modeling annotated data,” *ACM SIGIR*’03.

[2] J. Jeon & R. Manmatha, “Using maximum entropy for automatic image annotation,” *Proc. of CIVR*’04.

[3] J. Jeon, et al., “Automatic image annotation and retrieval using cross-media relevance models,” *ACM SIGIR*’03.

[4] J.-R. Bellegarda, “Exploiting latent semantic information in statistical language modeling,” *Proc. of the IEEE*, Vol.88, No.8, pp.1279-1296, 2000.

[5] K. Barnard, et al., “Matching words and pictures,” *Journal of Machine Learning Research*, pp. 1107-1135, Vol. 3, 2003.

[6] P. Carbonetto, et al., “A statistical model for general contextual object recognition”, *Proc. Of ECCV*’04.

[7] P. Duygulu, et al., “Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary,” *Proc. of ECCV*’02.

[8] Ricardo B. Y. & Berthier R.-N., *Modern Information Retrieval*, Addison Wesley, 1999.

[9] S. Gao, W. Wu, C.-H. Lee and T.-S. Chua, “A MFoM learning approach to robust multiclass multi-label text categorization,” *Proc. of ICML*’04.

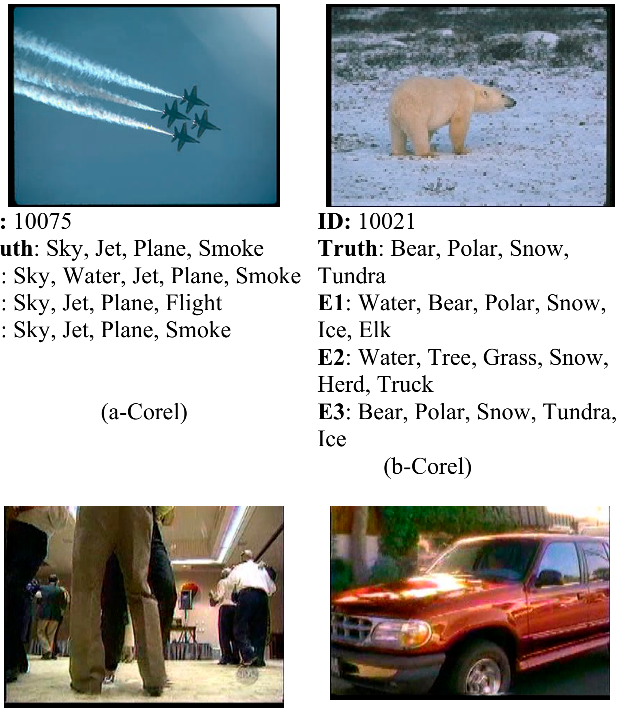
[10] S. Kumar & M. Hebert, “Discriminative fields for modeling spatial dependencies in natural images,” *Proc. of NIPS*’04.

[11] S.-L. Feng, et al., “Multiple Bernoulli relevance models for image and video annotation,” *Proc. of CVPR*’04.

[12] V. Lavrenko, et al., “A model for learning the semantics of pictures,” *Proc. of NIPS*’03.

[13] Wang D.-H., et al., “Discriminative fusion approach for automatic image annotation,” *Proc. of MMSP*’05.

[14] X.-M. He, et al., “Multiscale conditional random fields for image labeling,” *CVPR*’04.



ID: 10075
Truth: Sky, Jet, Plane, Smoke
E1: Sky, Water, Jet, Plane, Smoke
E2: Sky, Jet, Plane, Flight
E5: Sky, Jet, Plane, Smoke

(a-Corel)

ID: 10021
Truth: Bear, Polar, Snow, Tundra
E1: Water, Bear, Polar, Snow, Ice, Elk
E2: Water, Tree, Grass, Snow, Herd, Truck
E3: Bear, Polar, Snow, Tundra, Ice

(b-Corel)

ID: 19980104_ABC_85
Truth: Non-studio_setting, People
E1: Non-studio_setting, People
E2: People
E3: Non-studio_setting, People

(c-TRECVID)

ID: 19980104_ABC_151
Truth: Car/Bus/Truck
E1: Car/Bus/Truck, People
E2: People
E3:Car/Bus/Truck,Outdoors, Road

(d-TRECVID)

Figure 2 Some annotation examples