BOOSTING OF MAXIMAL FIGURE OF MERIT CLASSIFIERS FOR AUTOMATIC IMAGE ANNOTATION

Filippo Vella, Chin-Hui Lee** and Salvatore Gaglio**

*Istituto di Calcolo e Reti ad Alte Prestazioni, I.C.A.R.-C.N.R., Sede di Palermo
**Electrical and Computer Engineering Department, Georgia Institute of Technology
vella@pa.icar.cnr.it, chl@ece.gatech.edu, gaglio@pa.icar.cnr.it

ABSTRACT

Visual information contained in a scene is very complex and can be represented with multiple features describing aspects of the entire information. In this paper we propose a boosting approach to automatic image annotation by building strong classifiers based on multiple collections of weak concept classifiers with each collection focused on a single visual feature. The weak classifiers are trained with a maximal figure-of-merit learning approach. By exploiting multiple features the boosting procedure allows to build classifiers able to pick the most discriminative feature for the specific annotation task.

Index Terms**—** Image Annotation, Text Categorization, Multi-Topic, Maximal Figure of Merit, Boosting

1. INTRODUCTION

Automatic Image Annotation (AIA) aims at associating image content with a set of pre-defined textual labels, also known as concepts or keywords. Many techniques have been proposed to characterize the joint distribution between the keywords and the visual content being represented in term of symbolic elements, known as visual terms. By doing so each image can be converted into a document vector in a similar way to what is done in a vector based representation of text documents in information retrieval [15]. Image classification can now be cast as a multi-topic text categorization problem [16] in which a set of topics, or class labels, are assigned to a test image. Such symbolic representation, or tokenization, is used in translation model (TM) [1][5], maximum entropy (ME) [11], Markov random field (MRF) [4] and conditional random field (CRF) [10].

In this study we propose an approach to automatic image annotation with boosting of a collection of weak classifiers. They are trained with a maximal figure-of-merit (MFoM) [8] learning framework designed to maximize any performance metric. It has been successfully applied to automatic image annotation [8].

In previous studies on AIA modeling images are often characterized with a single global set of visual terms. In this paper each image is represented by multiple sets of visual terms, each is generated from a different visual feature. One dedicated set of MFoM-trained classifiers can therefore be obtained for each particular feature. These MFoM-trained classifiers are considered as *weak* classifiers and a boosting procedure is applied to build a global set of *strong* classifiers. Since AIA is cast as a multi-topic classification problem the conventional multiclass AdaBoost.M2 [7] could be applied directly but for the weaker requirements here a set of parallel AdaBoost.M1 is used instead.

An example of boosting of classifiers based on linear discriminant analysis (LDA) has been proposed [14] [17]. In [14] a particular form of LDA is applied to the specific task of face recognition. Boosting is also shown to improve classification performance [17] in the case the target metric is fixed. In this paper we propose a boosting framework that is applicable to any desirable figure-of-merit.

The remainder of the paper is organized as follow: Section 2 discusses the representation of the image in vector spaces characterized by multiple visual terms. Section 3 describes the application of MFoM to train AIA models. Section 4 deals with the boosting process of MFoM-based classifiers. Section 5 shows experimental results. Finally we summarize our findings in Section 6.

2. VECTOR-BASED IMAGE REPRESENTATION

Image content is typically very rich and analyzed scenes are usually full of clattered elements. Information captured in a generic picture has a number of multiple components that human visual system is able to catch as significant elements in a scene. Different representations are conceivable for a multi-purpose characterization of images. A different set of features is often selected according to the aspect of information that is more relevant to a particular task. Starting from the values of the visual features, a symbolic representation of images can be accomplished by extracting a set of fundamental units, or tokens, for all images.

1-4244-1437-7/07/$20.00 ©2007 IEEE

2.1. Feature Symbolic Level

Visual features, such as color, texture and shape, are often extracted from an image to describe its content. The distribution of the feature vectors is usually not evenly distributed, and tends to have different mass concentrations in different parts of the vector space. The set of points of high concentration, often called *visual terms*, can be used as a basis to represent the generic visual information.

This approach allows visual terms to emerge from a data set and build a generic set of symbols that is limited only by the coverage of the images training set. In the same way the discriminative power of the visual term for a particular feature is related to its characteristic and the correlation with the semantic content of images.

Although *k-means* algorithms have been widely used in AIA to extract a set of tokens [1][5][11][12], in this work the extraction of the visual terms has been achieved through vector quantization using the *LBG* algorithm [13].

2.2. Image representation

Consider a set of visual terms, $A=\{A_1,A_2,...,A_M\}$, with each element describing a specific visual characteristic. An image can be represented by a vector, $V=(v_1,v_2,...,v_M)$ where the *i-th* component takes into account occurrence count of the term A_i in the image. Obviously the complexity of the visual information is captured more reliably if more characteristics are used. A way to integrate information coming from heterogeneous features is to consider a unique composite vector and extract a unique visual vocabulary (set of visual terms) from it. This solution, although largely used, has drawbacks due to the computation cost of extracting a base for vector as long as the sum of all the features dimensions, and has to repeat the entire process from scratch each time a new feature is added to the previous ones. An interesting alternative is the usage of codebooks from different features and putting together heterogeneous visual information at the symbolic level.

To improve the representation capability of each visual dictionary, each single codebook can be exploited in a more extensive manner by representing the visual content in terms of both unigrams and spatially displaced bigrams[9]. The increased expressivity of the bigrams often outperforms, at the cost of an increased dimensionality, the results achieved with the mere application of unigrams. Considering a codebook for a single feature of M elements, the total dimension of the image vector is, in this case, $M*M+M$. For a codebook of 64 elements the total vector dimension is 4160, and it is increased to 16152 when M=128. To enhance the indexing power of each element, a normalized entropy of the representation for both the unigrams and bigram, can be computed [2]

3. AUTOMATIC IMAGE ANNOTATION WITH MFoM-TRAINED CLASSIFIERS

In AIA the training image set is given as a collection of pairs formed by a D-dimensional vector of values, representing the image, and one or more manually assigned 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.

In this study we used LDF-based classifiers to model the concept space because it is easy to separate high-dimension vector spaces with simple LDF models. MFoM learning can be applied to optimizing any desired figure-of-merit. It was shown to give good AIA performance [8] when a single visual dictionary is considered.

The classifier is formed by a set of functions $g_j(X,\Lambda_j)$ that are equal in number to the cardinality of the keywords set and for each of them the set of the parameters Λ_j is trained in order to discriminate the positive from negative samples. In the annotation stage, multiple relevant keywords are assigned to an image X, according to the following multiple-label decision rule,

$$\begin{cases} \text{Accept } X \in C_j \text{ if } g_j(X;\Lambda_j) - g_j^-(X;\Lambda^-) > 0 & \\ & 1\leq j\leq N \\ \text{Reject } X \in C_j, \textit{ Otherwise} & \end{cases} \qquad \textbf{(1)}$$

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

$$g_j^-(\mathrm{X};\Lambda^-) = \log\left[\frac{1}{\left|C_j^-\right|}\sum_{i\in C_j^-}\exp(g_i(\mathrm{X};\Lambda_i))^\eta\right]^{1/\eta} \qquad \textbf{(2)}$$

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. (2) measures the competing score as a geometric average of scores of all competing categories and works 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. (1) is high. Fortunately, the nature of multi-category learning makes it possible to compare scores from N concept models. In this way a comparison with the top-M keyword candidates gives a reliable evaluation for the task.

3.1. MC MFoM Learning

In MC MFoM learning, the parameter set $\Lambda = \{\Lambda_j, 1\leq j\leq N\}$ is estimated by optimizing a metric-oriented objective function which is continuous and differentiable, and is specially designed for approximating any performance metric, e.g. precision, recall or F1, based on error counts.

For the classifier, a linear discriminant classifier (named LDU or g-unit), apt to discriminate positive from negative samples, is trained for each keyword. The LDF has the form:

$$g_j(X,\Lambda_j) = W_j \cdot X + b_j \qquad \textbf{(3)}$$

where W_j and b_j are the parameters for the j-th concept model. A direct measure of misclassification is defined in terms of the score and anti-discriminant function as:

$$d_j(X;\Lambda) = -g_j(X,\Lambda) + g_j^-(\mathrm{X},\Lambda^-) \qquad \textbf{(4)}$$

If eq. (4) is negative image X is labeled with the j-th label while if it is positive competing labels are assessed as more adherent to the input. With the above definitions, most commonly used metrics (e.g. precision, recall and F1) can be defined in terms of d_j functions.

In the following series of experiments an objective function considering both false negative and false positive errors is considered. Directly derived from the Det curves, a det error is defined as:

$$DetE = \sum_{1\le j\le N} \frac{FP_j + FN_j}{2\cdot N} \qquad \textbf{(5)}$$

where FP_j and FN_j are the number of false positive and the false negative samples for the j-th label .

The linear discriminant functions are trained by minimizing the Det Error with a generalized probabilistic descent algorithm [8].

4. BOOSTING OF MFoM

Starting from a description of the training images in terms of different visual features, a multiple description for each image is available. Instead of considering a single visual dictionary with combined features, a different visual dictionary is extracted for each feature with a separate set of MFoM classifiers trained to minimize of the Det Error.

Given a specific visual feature, each g-unit is trained to minimize the Det error. The set of all the g-units that are in number equal to the number of labels for the number of the MFoM classifiers (or visual dictionaries), is then considered as the set of the *weak* classifiers, h_t .

Each g-unit is oriented to the detection of a single label or keyword and to make all the g-units comparable, a threshold to each output is imposed. Due to the characteristic of the classifier, each unit will have better performance when detecting its own class and worse performance when images of other classed are presented as input.

To build a system able to detect the presence of different labels a strong classifier, built through the boosting algorithm[7], is computed for each label. Considering that the g-unit values are the same for all the classes, the overhead for extending the classifier to all the classes is limited to the application of the boosting algorithm for each target class.

The description of the algorithm is referred to the boosting for a single label but for all the other labels the process can be replicated without a significant modification.

Given a training set, (x_1,y_1),…, (x_n,y_n), where x_i is a representation of the generic image and y_i indicates the presence or the absence of the specific label, the distribution D_t is computed as shown in the following figure:

```
Initialize weights $w_{1,i}$
for $t$=1,…,$T$
    1. normalize weights $w_{t,i}$
    2. evaluate the weighted error for each weak
       classifier
           $\varepsilon_j = \sum_i w_i \left| h_j(x_i) - y_i \right|$
    3. choose the $h_t$ with the lowest $\varepsilon_t$
    4. update the weights:
           $w_{t+1,i} = w_{t,i}\beta_t^{1-e_i}$
       where $e_i$=0 if $x_i$ is classified correctly, 1 otherwise
       and $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$
Final classifier:
    $H_{final}(x) = \begin{cases} 1 & \sum_{t=1}^{T} \alpha_t h_t(x) \ge \frac{1}{2}\sum_{t=1}^{T} \alpha_t \\ 0 & otherwise \end{cases}$
where $\alpha_t = \log\frac{1}{\beta_t}$
```

Figure 1, The Ada boost algorithm applied to MFoM

The boosted classifiers, built by considering the output of limiting the g-units with some thresholds, allows us to select the most reliable detector for each label and to weight their outputs according to their discrimination power for the target class.

5. EXPERIMENTAL RESULTS

The proposed approach is tested on the COREL data set consisting of 5000 images divided in 50 classes. 4500 images are used for training while the remaining 500 images are used for testing. For analysis purposes each image is partitioned into a collection of 16x16 macro-blocks and characterized by color, represented in the color spaces as RGB, YUV, and Lab, and texture, represented as energy distribution of Gabor wavelet (GAB) and fast Fourier transform (FFT). For each visual feature a set of 128 visual terms are selected and images are represented considering both unigrams and spatially displaced bigrams. A set of MFoM-trained classifiers are obtained for each feature and the values of the trained g-units is considered as input for the boosting procedure.

In Table 1 we show the performance of boosting when multiple visual dictionaries are used. The number of stages for the strong classifier is set to the half of the available weak classifiers. It clearly shows performance improvement in terms F1 measures and reduction of Det Error when multiple visual features are used to characterize images.

	RGB, GAB	RGB,GAB FFT	RGB,GAB FFT,YUV	RGB,GAB FFT,YUV LAB
F1	0.20	0.22	0.27	0.32
DetE	0.28	0.28	0.26	0.23

Table 1, Precision, Recall and F1 measure for Boosting of MFoM involving different visual dictionaries

In Figure 2 are depicted the Precision versus Recall curves for the same experimental set presented in Table 1. The figure shows the increment in precision for equal values of recall when multiple features are used to train the MFoM classifiers.

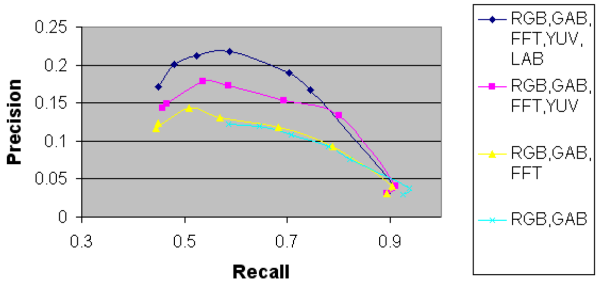

Figure 2, Precision vs recall graph when multiple features are used

In Table 2 a comparison with the state-of-art techniques in AIA is presented. The values of precision and recall for boosting of MFoM are compared with the published results of TM [4], CMRM [12], ME [11] and MBRM [6]. It must be said that although the table shows better performance for the proposed method, a direct comparison is impossible due to the fact that different features were used to describe the visual content, and different number of annotation labels were used in different models.

	TM	CMRM	ME	MBRM	**Boost MFoM**
Prec	0.06	0.10	0.09	0.24	0.22
Recall	0.04	0.09	0.12	0.25	0.59
F1	0.05	0.09	0.10	0.24	0.32

Table 2, Comparison of Precision and Recall with the state of art image annotation techniques

6. CONCLUSION

We propose a boosting approach to enhance multiple sets of MFoM-trained classifiers, each is trained with a different visual dictionary. The resulted global strong classifier allows the exploitation of heterogeneous features characterizing input image in a straightforward and modular manner. Experimental results show good performance with the proposed approach when compared with published results from state-of-art AIA models. As future work, the comparison with information fusion techniques applied at the same input information can corroborate the effectiveness of the proposed approach.

7. REFERENCES

[1] K. Barnard, et al., "Matching words and pictures", *Journal of Machine Learning Research*, Vol.3, pp1107-1135, 2003.

[2] J.-R. Bellegarda, "Exploiting latent semantic information in statistical language modelling", *Proc. of the IEEE*, Vol.88, No.8, pp1279-1296, 2000.

[3] D. Blei, M.-I. Jordan, "Modeling annotated data", *ACM SIGIR*, 2003

[4] P. Carbonetto, et al., "A statistical model for general contextual object recognition", *Proc. of ECCV*, 2004.

[5] P. Duygulu, et al., "Object recognition as machine translation: Learning a lexicon for a fixed vocabulary", *Proc. of ECCV*, 2002.

[6] S.-L. Feng, R. Manmatha and V. Lavrenko, "Multiple Bernoulli relevance models for images and video annotation", *Proc. of ICML*, 2004

[7] Y. Freund and R.E. Schapire. A decision-theoretic generalization of online learning and an application to boosting. In Computational Learning, Theory (Eurocolt), 1995.

[8] 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*, 2004.

[9] S. Gao, D.-H. Wang and C.-H. Lee, "Automatic Image Annotation through Multi-Topic Text Categorization", *Proc. of ICASSP*, 2006.

[10] X.-M. He et al., "Multiscale conditional random fields for automatic image annotation", *Proc. of CVPR*, 2004

[11] J. Jeon, R. Manmatha, "Using maxmum entropy for automatic image annotation", *Proc. of CIVR*, 2004.

[12] J. Jeon, R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models", *ACM SIGIR* 2003

[13] Y. Linde, A. Buzo, R. Gray, An Algorithm for Vector Quantizer Design. *IEEE Transaction on Communications*, vol. 28 no. 1, pp 84–94, 1980.

[14] J. Lu, K-N. Plataniotis, A.N. Venetsanopoulos, Boosting Linear Discriminant Analysis for face recognition, *Proc of Int. Conf. on Image Processing*, 2003.

[15] G. Salton, *The SMART Retrieval System*, Prentice-Hall, Englewood Cliffs, 1971

[16] F. Sebastiani, Machine Learning in Automated Text Categorization, *ACM Computer Surveys*, vol. 34, no. 1, 2002

[17] M. Skurichina, R.P. Duin, Boosting in Linear Discriminant Analysis, *LNCS Proceedings of the First International Workshop on Multiple Classifier Systems,* 2000