TELECOM ParisTech at ImageCLEF 2010 Photo Annotation Task: Combining Tags and Visual Features for Learning-Based Image Annotation

Hichem Sahbi * Xi Li *,**

*CNRS LTCI, UMR 5141
TELECOM ParisTech
46, rue Barrault, 75634 Paris Cedex, France
**NLPR, CASIA, Beijing, China
hichem.sahbi@telecom-paristech.fr
lixichinanlpr@gmail.com

Abstract. In this paper, we describe the participation of TELECOM ParisTech in the ImageCLEF 2010 Photo Annotation challenge. This edition focuses on promoting combination between visual and tag features in order to enhance photo annotation. An image collection is supplied with tags which are used both for training and testing. Our training approach consists of building SVM classifiers and kernels which take into account the similarity between visual features as well as tags. The results clearly corroborate (i) the complementarity of tags and visual descriptors and (ii) the effectiveness of SVM classifiers in photo annotation.

1 Introduction

Recent years have witnessed a rapid increase of image sharing spaces, such as Flickr, due to the spread of digital cameras and mobile devices. An urgent need is how to effectively search these huge amounts of data and how to exploit the structure of these sharing spaces. A possible solution is CBIR (Content-Based Image Retrieval); where images are represented using low-level visual features (color, texture, shape, etc.) and searched by analyzing and comparing those features. However, low-level visual features are usually unable to deliver satisfactory semantics, resulting in a gap between them and the high-level human interpretations. To address this problem, a variety of machine learning techniques were introduced in order to discover the intrinsic correspondence between visual features and semantics of images and allow to predict keywords for images.

2 Related Work

Conventionally, image annotation is converted into a classification problem. Existing state of the art methods (for instance [1, 2]) treat each keyword or concept as an independent class, and then train the corresponding concept-specific classifier to identify images belonging to that class, using a variety of machine learning techniques such as hidden Markov models [2], latent Dirichlet allocation [3], probabilistic latent semantic analysis [4], and support vector machines [5]. The aforementioned annotation methods may also be categorized into two branches; region-based requiring a preliminary step of image segmentation [2, 12], and holistic [6, 25] operating directly on the whole image space. In both cases, training is achieved in order to learn how to attach keywords with the corresponding visual features.

The above annotation methods heavily rely on their visual features for image annotation. Due to the semantic gap, they are unable to fully explore the semantic information inside images. Another class of annotation methods has emerged that takes advantage of extra information (tags, context, users' feedback, ontologies, etc.) in order to capture the correlations between images and concepts. A representative work is the cross-media relevance model (CMRM) [6, 9], which learns joint statistics of visual and concepts and its variants [7, 8]. The model uses the keywords shared by similar images to annotate new ones. In [22], the similarity measure between images integrates contextual information for concept propagation. Semi-supervised annotation techniques were also studied and usually rely on graph inference [10–13]. The original work, in [3, 26], is inspired from machine translation and considers images and keywords as two different languages; in that case, image annotation is achieved by translating visual words into keywords.

Other existing annotation methods focus on how to define an effective distance measure for exploring the semantic relationships between concepts in large scale databases. In [19], the Normalized Google similarity Distance (NGD) is proposed by exploring the textual information available on the web. It is a measure of semantic correlations derived from counts returned by Google's search engine for a given set of keywords. Following the idea of [19], the Flickr distance [20] is proposed to precisely characterize the visual relationships between concepts. Each one is represented by a visual language model in order to capture its underlying visual characteristics. Then, a Flickr distance is defined, between two concepts, as the square root of Jensen-Shannon (JS) divergence between the corresponding visual language models. Other techniques consider extra knowledge derived from ontologies (such as the popular WordNet [14–16]) in order to enrich annotations [21]. The method in [14] introduces a visual vocabulary in order to improve translation model in the preprocessing stage of visual feature extraction. A directed acyclic graph is used to model the causal strength between concepts, and image annotation is performed by inference on this graph [15]. In [17, 18], the semantic ontology information is integrated in the post processing stage in order to further refine initial annotations.

3 Motivation and The Proposed Method at a Glance

Among the most successful annotation methods, those based on machine learning and mainly support vector machines; show a particular interest as they are performant and theoretically well grounded [24]. Support vector machines [23], basically require the design of similarity measures, also referred to as *kernels*, which should provide high values when two images share similar structures/appearances and should be invariant, as much as possible, to the linear and non-linear transformations. They also satisfy positive definiteness which ensures, according to Vapnik's SVM theory [24], optimal generalization performance and also the uniqueness of the SVM solution. In practice, kernels should not depend only on intrinsic aspects of images (as images with the same semantic may have different visual and textual features), but also on different sources of knowledge including context.

In this work, we introduce an image annotation framework based on a new similarity measure which takes high values not only when images share the same visual content but also the same context. The context of an image is defined as the set of images, with the same tags, and exhibiting better semantic descriptions, compared to both pure visual and tag based descriptions. The issue of combining context and visual content for image retrieval is not new (see for instance [28–30]) but the novel part of this work aims to (i) integrate context, in similarity design useful for classification and annotation, and (ii) plug this similarity in support vector machines in order to take benefit from their well established generalization power [24]. This type of similarity will be referred to as context-based while those relying only on the intrinsic visual or textual content will be referred to as context-free. Again, our proposed method goes beyond the naive use of low level features and context-free similarities (established as the standard baseline in image retrieval) in order to design a similarity applicable to annotation and suitable to integrate the "contextual" information taken from tagged datasets. In the proposed method, two images (even with different visual content and even sharing different tags) will be declared as similar if they share the same visual context[1]. This is usually useful as tags in data may be noisy and misspelled. Furthermore, the intrinsic visual content of images might not always be relevant especially for categories exhibiting large variation of the underlying visual aspects.

Through this work, an image database is modeled as a graph where nodes are pictures and edges correspond to shared tags (links) between images. We design our similarity as the solution of a constrained energy function containing a fidelity term which measures visual similarity between images and a context criterion that captures the similarity between the underlying links.

[1] Visual context is defined as the set of images sharing the same tags.

4 Evaluation

4.1 MIR Flickr/ImageCLEF Collection

We evaluated our annotation method on the MIR Flickr dataset containing $18,000$ images belonging to 93 categories (for instance "sky, clouds, water, sea, river,..."), among them $8,000$ are used for training and $10,000$ for testing. The whole dataset is annotated but ground truth is provided only for the training set. The MIR Flickr collection contains $1,386$ tags (provided by the Flickr users) which occur in at least 20 images, with an average total number of 8.94 tags per image (see Fig. 1 and [32]).

Fig. 1. This figure shows samples of images taken from the ImageCLEF 2010 Photo Annotation Task Database.

4.2 Indexing and Annotation

Recent years have witnessed a great success of the bag-of-features representation in a wide range of application, such as image retrieval, image classification, image segmentation, object recognition, etc. Inspired by text classification, visual feature spaces are conventionally partitioned by vector quantization (e.g. kmeans) into several subspaces, each of which corresponds to a visual word. As a consequence, the bag-of-feature representation is converted to the bag-of-words (BoW). Since using a basic histogram of orderless visual words, the BoW representation only reflects the global statistical properties of visual words, and ignores their spatial layout. Therefore, the orderless BoW representation has a low descriptive capability of capturing the geometric relationships among visual words. Motivated by this, we use, in this evaluation campaign, the same approach as in [27] in order to better capture the spatial layout of images. The algorithm is based on a spatial pyramid representation, which constructs a multi-level spatial pyramid by block division. For any block at each level, a traditional BoW representation in the SIFT feature space is used. In this way, we have a set of block-specific BoW histograms at multiple levels. As a result, the geometric relationships among visual words can be effectively captured.

Given a test picture, the goal is to predict which categories (object classes) are present into that picture. This task is commonly known as concept detection. For this purpose, we trained "one-versus-all" SVM classifiers for each category; we repeat this training process through different folds (20 times), for each category, and we take the average score of the underlying SVM classifiers on the test picture. This makes classification results less sensitive to sampling and unbalanced classes. Performances are reported using the Mean Average Precision (MAP), the Equal Error Rate (EER) and the Area Under Curve (AUC). Higher MAP, AUC and lower EER imply better performance. Figs. (2, 3, 4) show the annotation results of our best ImageCLEF run through different classes.

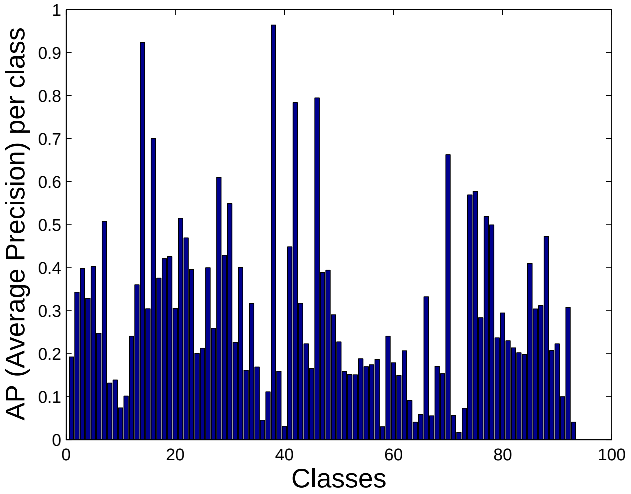

Fig. 2. This figure shows the average precision per class.

5 Conclusion

We introduced in this work our participation in the ImageCLEF 2010 Photo Annotation Task. Our annotation method takes into account image features as well as their context links (taken from tags in the MIR Flickr collection) in order to achieve SVM learning and classification. Future extensions of this work include extra processing of these tags prior to SVM learning and further evaluations in the next campaigns.

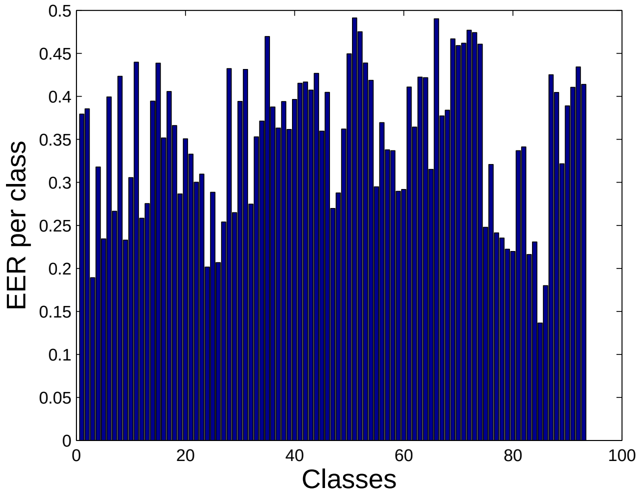

Fig. 3. This figure shows the Equal Error Rate per class.

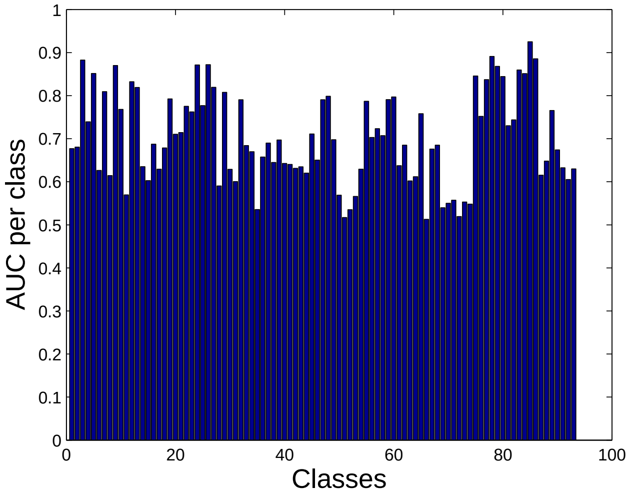

Fig. 4. This figure shows the Area Under Curve per class.

Acknowledgement

This work is supported by the French National Research Agency (ANR) under the AVEIR project.

References

1. G. Carneiro and N. Vasconcelos, “Formulating semantic image annotation as a supervised learning problem”, in *Proc. of CVPR,* 2005.
2. J. Li and J. Z. Wang, “Automatic linguistic indexing of pictures by a statistical modeling approach,” *IEEE Trans. on PAMI.,* 25(9):1075-1088, 2003.
3. K. Barnard, P.Duygululu, D. Forsyth, D. Blei, and M. Jordan, “Matching words and pictures,” *The Journal of Machine Learning Research*, 2003.
4. F. Monay and D. GaticaPerez, “PLSA-based Image AutoAnnotation: Constraining the Latent Space,” in *Proc. of ACM International Conference on Multimedia*, 2004.
5. Y. Gao, J. Fan, X. Xue, and R. Jain, “Automatic Image Annotation by Incorporating Feature Hierarchy and Boosting to Scale up SVM Classifiers,” in *Proc. of ACM MULTIMEDIA*, 2006.
6. J. Jeon, V. Lavrenko, and R.Manmatha, “Automatic image annotation and retrieval using cross-media relevance models,” in *Proc. of ACM SIGIR*, pp. 119-126, 2003.
7. V. Lavrenko, R. Manmatha, and J. Jeon, “A model for learning the semantics of pictures,” in *Proc. of NIPS*, 2004.
8. S. Feng, R. Manmatha, and V. Lavrenko, “Multiple Bernoulli relevance models for image and video annotation,” in *Proc. of ICCV*, pp. 1002-1009, 2004.
9. J.Liu, B.Wang, M.Li, Z.Li, W.Ma, H.Lu, and S.Ma, “Dual cross-media relevance model for image annotation,” in *Proc. of ACM MULTIMEDIA*, pp. 605-614, 2007.
10. X. Wan, J. Yang, and J. Xiao, “Manifold-ranking based topic-focused multi-document summarization,” in *Proc. of IJCAI*, pp. 2903-2908, 2007.
11. D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf. Ranking on data manifolds, in *Proc. of NIPS*, 2004.
12. J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma, ”Image annotation via graph learning,” *Pattern Recognition*, 42(2):218C228, 2009.
13. J. Liu, M. Li, W. Ma, Q. Liu, and H. Lu, “An adaptive graph model for automatic image annotation,” in *Proc. of ACM International Workshop on Multimedia Information Retrieval*, pp. 61C70, 2006.
14. M. Srikanth, J. Varner, M. Bowden, and D. Moldovan, “Exploiting ontologies for automatic image annotation,” in *Proc. of SIGIR*, pp. 552-558, 2005
15. Y. Wu, E. Y. Chang, and B. L. Tseng. “Multimodal metadata fusion using causal strength,” in *Proc. of ACM MULTIMEDIA*, pp. 872-881, 2005.
16. G. A. Miller, “Wordnet: a lexical database for English,” Commun. ACM, 38(11):39-41, 1995.
17. C. Wang, F. Jing, L. Zhang, and H. J. Zhang, “Image annotation refinement using random walk with restarts,” in *Proc. of ACM MULTIMEDIA*, pp. 647-650, 2006.
18. Y. Jin, L. Khan, L. Wang, and M. Awad, “Image annotations by combining multiple evidence & wordNet,” in *Proc. of ACM MULTIMEDIA*, pp. 706-715, 2005

19. R. Cilibrasi and P. M. B. Vitanyi, “The google similarity distance,” *IEEE Transactions on Knowledge and Data Engineering*, 2007.

20. L. Wu, X. Hua, N. Yu, W. Ma, and S. Li, “Flickr distance,”, in *Proc. of ACM MULTIMEDIA*, 2008.

21. Y. Wang and S. Gong, “Translating Topics to Words for Image Annotation,” in *Proc. of ACM CIKM*, 2007.

22. Zhiwu Lu, Horace H.S. Ip, and Q. He, “Context-Based Multi-Label Image Annotation,” in *Proc. of ACM CIVR*, 2009.

23. Boser B, Guyon I., and Vapnik V, ” An training algorithm for optimal margin classifiers” in *In Fifth Annual ACM Workshop on Computational Learning Theory, Pittsburgh*,1992.

24. V. Vapnik, ”Statistical Learning Theory.”, in *A Wiley-Interscience Publication”*, 1998”.

25. C. Wang, S. Yan, L. Zhang, H. Zhang, “Multi-Label Sparse Coding for Automatic Image Annotation,”, in *Proc. of CVPR,* 2009.

26. P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth, “Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary,” in *Proc. of ECCV*, 2002.

27. S. Lazebnik, C. Schmid, and J. Ponce, “Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories”, in *Proc. of CVPR,* 2006.

28. A.C. Gallagher, C.G. Neustaedter, L. Cao, J. Luo, and T. Chen, ”Image Annotation Using Personal Calendars as Context”, in *Proc. of ACM Multimedia”*,,2008

29. Cao L., Luo J. and Huang T.S., ”Annotating Photo Collection by Label Propagation According to Multiple Similarity Cues”, in *Proc. of ACM Multimedia”,* 2008

30. Y.H. Yang, P.T. Wu, C.W. Lee, K.H Lin, W.H. Hsu, and H. Chen , “ContextSeer: Context Search and Recommendation at Query Time for Shared Consumer Photos,” in *Proc. of ACM Multimedia”,* 2008

31. D. Haussler, “Convolution Kernels on Discrete Structures,” in *Technical Report UCSC-CRL-99-10, University of California in Santa Cruz, Computer Science Department, July*, 1999

32. S. Nowak and Mark Huiskes, “New Strategies for Image Annotation: Overview of the Photo Annotation Task at ImageCLEF 2010,” in *The Working Notes of CLEF 2010, Padova, Italy*, 2010.