TELECOM ParisTech at ImageClef 2009: Large Scale Visual Concept Detection and Annotation Task

Marin Ferecatu[*†] Hichem Sahbi[†*]
[*] Institut TELECOM, TELECOM ParisTech
[†] CNRS LTCI, UMR 5141
46, rue Barrault, 75634 Paris Cedex, France
Marin.Ferecatu@telecom-paristech.fr
Hichem.Sahbi@telecom-paristech.fr

Abstract

In this paper we describe the participation of TELECOM ParisTech in the Large Scale Visual Concept Detection and Annotation Task at the ImageClef 2009 challenge. This year, the focus was in the extension of (i) the amount of data available for training and testing, and (ii) the number of concepts to be annotated. We use Canonical Correlation Analysis in order to infer a latent space where text and visual description are highly correlated. Starting from a visual description of a test image, we first map it into the latent space, then we predict the underlying text features (and also annotations) which best fit the visual ones in the latent space. Our method is very fast while achieving good results.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation.

Keywords

Image annotation, Canonical Correlation Analysis, Text and image descriptors.

1 Task description

The Large Scale Visual Concept Detection and Annotation Task (referred here as VCDAT) offers a unified framework for image annotation to the participating teams: the goal is to annotate each test image with keywords describing the visual content and its semantic interpretation. The task provides annotated images using 53 concepts; all images have multiple annotations and the concepts are organized in a small ontology. The participants are allowed to use the relations between concepts for solving the annotation task. The training set consists of 5.000 images annotated with the 53 visual concepts and the test data consists of 13.000 photos. The participants are allowed to use only the training data in order to tune their algorithms.

Two evaluation measures are proposed: (a) per concept: false positive and false negative rates and (b) per image: a hierarchical measure that considers partial matches and calculates misclassification costs for each missing or wrongly annotated concept, based on structure information (distance between concepts in the hierarchy) and relationships from the ontology [13].

2 Summary of our approach

This year, the VCDAT focuses on scaling annotation algorithms to thousands of images and possibly more, which is indeed a very difficult task. Image annotation is still an unsolved problem and recent state of the art algorithms perform less than satisfactorily on most image databases [2, 5]. Image annotation is one branch of computer vision related to object detection and recognition; its goal is to decide whether an image contains one or multiple targeted objects and if yes, finds their *locations*. This problem is well studied and reasonably well solved for particular objects such as faces [18, 17] but remains reputedly difficult for many other classes of objects [10, 15].

Generally, local approaches, for instance those relying on keypoint extraction or image segmentation, are likely to offer better results, but at the expense of a much higher computational effort [12, 14]. Regardless the computational issues, VCDAT uses 53 concepts and many of them are *holistic*[1] so local (and also object based) methods are unlikely to provide descent results for this particular level of difficulty. Furthermore, local approaches hit the exterme variability of objects (concepts) into scenes and the limited amount of training images in order to capture this variability.

Instead, we focus on global approaches i.e., those which extract global image descriptions and easily handle large scale databases and annotations. This scalability will be achieved at the detriment of slight decrease of precision. Moreover, as we shall see, adding and training our system with new concepts is straightforward and does not require separate models for each one.

The remainder of this paper is organized as follows, we first describe our visual image and text features (see §3), then we discuss the application of Canonical Correlation Analysis (CCA) in order to infer a latent space where the two underlying representations are highly correlated (§4.) Given a visual description of a new (test) image, we first project it into the CCA latent space, then we infer text features as a linear combination of basic concepts which correlate the best with the visual one. Finally, we back-project the resulting text features into the (input) concept space and we normalize the projection coefficients between 0 and 1. A value close to 1 means that the corresponding concept is likely to be present into an image while a value close to 0 corresponds to an unlikely concept.

3 Text and visual content description

Visual descriptors. Global image descriptors have some properties that are very desirable in our case: (a) they have small memory footprint and thus fit into standard PCs without any specific storage requirements; (b) they are very fast to compute as they involve simple distance computation operations, guaranteeing real time responses; and (3) they do not include any a priori object model and thus can be applied to any target category. Indeed, global descriptors have been shown to perform well in this framework, for example with machine learning and data mining algorithms [2, 9, 4].

More precisely, we use a combination of color, texture and shape features, as follows. To represent color we use *weighted color histograms*: they provide a summary description of the color information including spatial measure in order to enphasize image regions that are interesting with respect to the visual content [16, 1]. As for *texture features* we use the power spectral density distribution in the complex plane. This has been shown to perform well when combined with color and shape histograms [11]. Roughly, a high energy spectrum concentrated at low frequencies highlights large scale informations in an image, while high frequencies correspond to textured regions (small scale details). In order to describe the *shape content* of an image we use standard edge orientation histograms. First, edges are extracted from images, then the gradient is computed using only the edge pixels. The orientation of the gradient is quantized w.r.t. the angle resulting into a histogram that is sensible to the general flow of lines in the image [8]. More details on image descriptors can be found in [3].

Text descriptors. We use the annotations provided for the training set in order to compute the text features. The latter have 53 dimensions, one for each concept c, indicating the presence or the absence of c. The resulting feature vector is very sparse; i.e., when applying principal component analysis (PCA), we found that 48 dimensions are sufficient in order to capture 100% of the statistical variance of the training data.

[1] Holistic means that the annotation is based on a global impression of a scene and not necessarily related to its physical objects.

4 Prediction using CCA

Canonical Correlation Analysis was first introduced by Hotelling [7] and it is used in order to capture linear relationships between two (or many) ordered[2] sample sets in different feature spaces. Canonical correlation analysis seeks a pair of linear transformations, one for each of the feature spaces, which map training and testing data into a common latent space. The latter is built in order to maximize the correlation between the sample sets in different feature spaces[6].

Given a test image, first we extract its visual feature vector and we project it into the CCA latent space. Then, we back-project the latent feature vector into the 53 dimensions of text space using the Moore-Penrose pseudo-inverse of the CCA transformation matrix. Now, annotations correspond to the entries among the 53 dimensions where the score is larger than a given threshold.

Training data consists of 5.000 images sharing 53 concepts. Fig. 1 shows the distribution of the number of images through different concepts. The most frequent one appears in 4656 images while the less frequent annotates only 18 images. Notice that both "very frequent" and "very rare" concepts are difficult to learn as the underlying positive and negative classes are clearly unblanced.

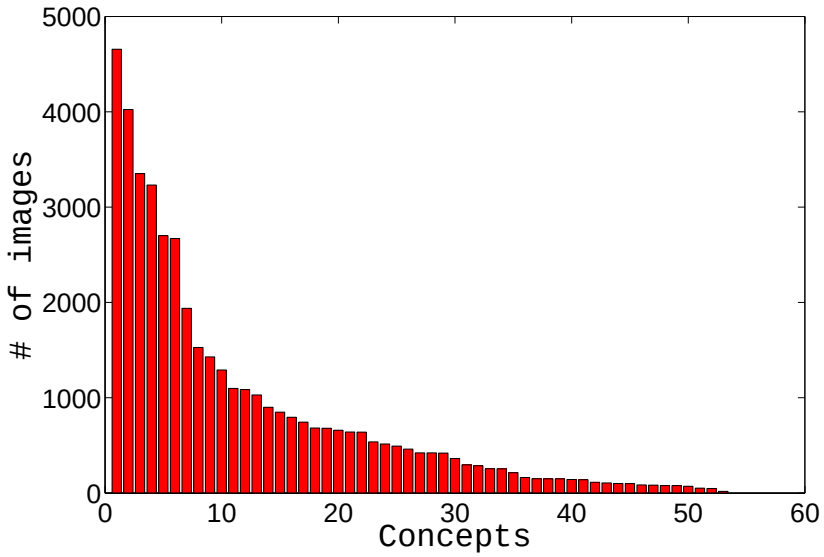

Figure 1: Number of images per concepts.

We randomly split the training set in two parts: one used for learning the CCA transform (4.000 images) and the other one used in order to evaluate the performance (1.000 images). Since the output of the algorithm has an asymptotic normal distribution, we normalize it to 0.5 mean and $1/6$ standard deviation. This ensures that 99.7 of the predicted scores lie between 0 and 1. Scores less than 0 (resp. larger than 1) are mapped to 0 (resp. 1).

The evaluation measure we use is the annotation error defined as the expected false negatives and false positives. For each concept c, we fix a threshold $\tau(c)$ and we annotate images with c if the underlying scores are larger that $\tau(c)$. Notice that $\tau(c)$ is fixed in order to minimize the error rate. We then linearly shift $\tau(c)$ to 0.5 in order to comply with the submission format.

On these challenging test images, our annotation method achieves relatively reasonable performances; the false positive error rate is 0.18 while the false negative one reaches 0.21. Nevertheless, our method is very efficient; in practice it tooks about a second in order to achieve training and prediction using a standard Pentium-M processor (with 2500 Mhz).

We also extended our method in order to use the ontology suggested by the challenge. Text features were enriched using this ontology in order to include all intermediate concepts and then propagate the

[2] One may define any arbitrary order for each sample set but should keep that order in different feature spaces.

annotation along the hypernyms tree. Notice that predictions include only the 53 concepts required by the benchmark. However, the ontology is too small in order to provide a noticeable improvement. Indeed, its total number of nodes is 68 where 53 (out of the 68) correspond to the candidate annotations. Again, we found that text features are still living into a subspace of 48 dimensions and this clearly shows that new extended concepts provide the same amount of information as initial ones.

5 Conclusion and perspectives

In this work we introduced the participation of TELECOM ParisTech in the Large Scale Visual Concept Detection and Annotation Task at ImageClef 2009. This year the task focuses in scalability of the annotation methods to large databases. Consequently, we use global, fast and easy to compute images descriptors that require very few computation resources. Our method constructs a latent space, using Canonical Correlation Analysis, where text and image features are highly correlated. It is extremely fast, it runs in less that a second both for training and for testing on a standard 2.5 GHz PC, and makes annotation effective and efficient in order to handle large scale databases.

Acknowledgements

This work was supported by the French National Research Agency (ANR) under the AVEIR[3] project, ANR-06-MDCA-002.

References

[1] Nozha Boujemaa, Julien Fauqueur, Marin Ferecatu, François Fleuret, Valérie Gouet, Bertrand Le Saux, and Hichem Sahbi. Ikona: Interactive generic and specific image retrieval. In *Proceedings of the International workshop on Multimedia Content-Based Indexing and Retrieval (MMCBIR'2001)*, 2001.

[2] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):5:1–60, 2008.

[3] Marin Ferecatu. *Image retrieval with active relevance feedback using both visual and keyword-based descriptors*. PhD thesis, INRIA—University of Versailles Saint Quentin-en-Yvelines, France, 2005.

[4] Theo Gevers and Arnold W. M. Smeulders. Content-based image retrieval: An overview. In G. Medioni and S. B. Kang, editors, *Emerging Topics in Computer Vision*. Prentice Hall, 2004.

[5] Allan Hanbury. A survey of methods for image annotation. *Journal of Visual Languages and Computing*, 19(5):617–627, 2008.

[6] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[7] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:312–377, 1936.

[8] A.K. Jain and A. Vailaya. Shape-based retrieval: a case study with trademark image databases. *Pattern Recognition*, 31(9):1369–1390, 1998.

[9] M. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State-of-the-art and challenges. *ACM Transactions on Multimedia Computing, Communication, and Applications*, 2(1):1–19, 2006.

[10] D. Lowe. A survey of methods for image annotation. *International Journal of Computer Vision*, 80(2), 2004.

[3] `http://aveir.lip6.fr`

[11] B.S. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley, 2002.

[12] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630, 2005.

[13] Stefanie Nowak and Hanna Lukashevich. Multilabel classification evaluation using ontology information. In *Proc. of the IRMLES Workshop*, 2009.

[14] Jean Ponce, Martial Hebert, Cordelia Schmid, and Andrew Zisserman. *Towards category-level object recognition*, volume 4170. Springer, 2006.

[15] A. Torralba, K. P. Murphy, , and W. T. Freeman. Sharing visual features for multiclass and multiview object detection. In *Proc. of CVPR*, 2004.

[16] Constantin Vertan and Nozha Boujemaa. Upgrading color distributions for image retrieval: can we do better? In *International Conference on Visual Information Systems (Visual2000)*, November 2000.

[17] P. Viola and M. Jones. Sharing visual features for multiclass and multiview object detection. In *Proc. of ICCV*, 2001.

[18] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *International Journal of Computer Vision*, 35(4):399 – 458, 2003.