IAM@ImageCLEFPhotoAnnotation 2009: Naïve application of a linear-algebraic semantic space

Jonathon S. Hare, Paul H. Lewis
Intelligence Agents Multimedia Group
School of Electronics and Computer Science, University of Southampton, Southampton, UK
{jsh2|phl}@ecs.soton.ac.uk

Abstract

This paper describes Southampton's submissions to the 2009 ImageCLEF photo annotation task. For the task we used an annotation system based on the idea of constructing semantic spaces, which was developed previously at Southampton. To represent the image content, we used a combination of different SIFT and Colour-SIFT features detected using the difference-of-Gaussian and MSER techniques. These features were converted into a visual term representation by applying vector quantisation using a codebook learnt from a hierarchical k-means clustering. In terms of EER and AUC, the annotator performs reasonably well, however, it struggles when evaluated using the hierarchical measure proposed for the task, due to the way the annotation confidences are thresholded.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance Evaluation*; I.4.9 [**Artificial Intelligence**]: Applications; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Automatic Annotation, Performance, Experimentation

Keywords

Image Content Analysis, Data Fusion, Semantic Space

1 Introduction

The ImageCLEF 2009 photo annotation task set the challenge of automatically annotating 13000 images with 53 annotation concepts. The allowable training data was limited to a set of 5000 images pre-labelled with the concepts. The images themselves were drawn from the MIR Flickr 25000 image collection [4]. Southampton's submissions to the task used a previously developed annotation system, with a combination of visual term features created from local descriptors of salient interest regions.

2 Methodology

As with many automatic annotation approaches, the methodology applied to this task involved extracting feature vectors for each of the images, and then feeding the features of a training set, together with annotations to a machine learning system. The machine learning system attempts to learn low-level relationships between all of the features and annotations. Once the training phase is complete, features from un-annotated images can be fed into the system to use the learnt relations to get predictions of annotations.

2.1 Visual Features

The images were represented by vectors of visual-term occurrences [11]. The visual-terms were created by finding interest points and extracting local feature descriptors, and then quantising to a pre-determined codebook. For the experiments in this task we used a combination of multi-scale difference-of-Gaussian interest regions with SIFT features [6], MSER regions [7] with SIFT features, and MSER regions with colour-SIFT features [1]. Each of the three region/feature combinations had its own 3125 term codebook created by applying hierarchical k-means [8] (5 levels with 5 clusters per node). The codebook size was not optimised in any way, and was chosen based on a best guess basis from previous experience with these feature morphologies and the machine learning technique described in the next subsection. The final image representation was created by appending the term-occurrence vectors from each of the region/feature representations to create a vector with 9375 dimensions.

2.2 Machine Learning

The machine-learning component is based on a linear-algebraic semantic space [3, 2], which is a development and generalisation of a text indexing technique called Cross-Language Latent Semantic Indexing [5]. This technique produces a vector space into which both visual-terms and keyword terms are mapped along with the images. Un-annotated images can then be projected into this space. Annotation was performed by projecting the test images into the space, and ranking the possible annotations based on their cosine similarity.

The use of the cosine similarity measure gives each possible annotation a score between -1 and 1; however, these scores are not themselves all that informative. A higher score does mean more confidence in an annotation, but only when considered against all the other annotations. For the purposes of evaluating the annotator using the EER and AUC measures this is not too much of a problem — we can just scale the scores to the 0..1 range (i.e add one and divide by two). However, as will be discussed in the next section, the hierarchical scoring measure [10] thresholds the annotation confidences at 0.5 to produce a binary indication of present/not-present. Unfortunately, for the semantic space this poses a big problem as the position of the threshold should ideally be set differently for each image, based on confidences of all the predicted annotations.

3 Experiments, Results and Discussion

We submitted three different runs to the task organisers. The first run was trained on the raw annotations. The second included a partial expansion of the annotation hierarchy [9] provided by the organisers, based on the non-abstract nodes (i.e. if Lake=true then water=true). The third included a full expansion of the hierarchy. The hierarchical expansion just means that a few extra annotation terms are fed into the machine learning component together with the leaf-node annotations already present for the image in question. The run titles and settings are shown in Table 1.

Run Title	Description
IAM Southampton_30_2_1245438072355.txt	Raw annotations
IAM Southampton_30_2_1245519187248.txt	Partial hierarchical expansion
IAM Southampton_30_2_1245519327555.txt	Full hierarchical expansion

Table 1: Description of submitted runs

Technique	EER	AUC
Mean	*0.373*	*0.553*
Median	*0.372*	*0.673*
Min	*0.234*	*0.070*
Max	*0.526*	*0.839*
Random	*0.500*	*0.499*
IAM Southampton_30_2_1245438072355.txt	*0.330*	*0.715*

Table 2: Summary of averaged EER and AUC scores over all annotation terms. The summary statistics were calculated using only for the best run of each participant.

3.1 Preliminary Results Analysis

The two runs that included the hierarchical information did not perform as well (based on average EER and AUC) as the one based on the raw annotations. Looking at the EER scores for each annotation term, the hierarchical methods were consistently worse performing. For these reasons the results for these runs will not be discussed further.

The EER and AUC scores are summarised in Table 2. Our scores are better than the averages of the other participants, however, they are still a fair way off of the top scores. Using a semantic space approach for annotation, from past experience, we expect that there will be a large diversity in the performance for different terms. This is due to differences in the amount of training data for each term, and also the amount of visual diversity that might be associated with a term. For example, visually specific annotation terms require less training data than diverse ones. Table 3 shows the top- and bottom-most 5 annotation terms. It is interesting that the worst performing terms are those that are rather general and unspecific. The top performing terms all have very specific visual representations.

Table 4 shows the results of our annotator using the hierarchical measure [10]. Unfortunately, because this scoring measure performs a binary thresholding operation on the confidence scores,

Annotation Term	EER
Sunset-Sunrise	0.232
Landscape-Nature	0.234
Night	0.237
Sea	0.243
Mountains	0.249
Aesthetic-Impression	0.416
Overall-Quality	0.436
Neutral-Illumination	0.466
Sports	0.470
Fancy	0.478

Table 3: Best and worst annotations by EER

the performance of our technique measures at the lower end of the spectrum of results from the different participants. As previously discussed, the semantic space annotation approach doesn't really permit the global setting of such a threshold.

Technique	**EER**	**AUC**
Mean	*0.684*	
Median	*0.752*	
Min	*0.390*	
Max	*0.829*	
Random	*0.384*	
IAM Southampton_30_2_1245438072355.txt	0.41897374	

Table 4: Summary of averaged hierarchical scores

3.2 Computational Performance and Implementation Details

The feature extraction phase was performed in parallel (4 images being processed at once) on a quad core machine (Intel Core 2 Quad @ 2.66Ghz, 8G ram, Redhat Enterprise 5.3). The time for image processing varied based on both the size of the image, and the image content. Timings for a typical image from the training set are shown in Table 5.

Training the semantic space took approximately 1 hour on a dual quad core 2.8GHz Xeon workstation running Mac OS X (the semantic space code is single threaded, so only uses a single core). We would estimate that no more than 1G of ram was used during the semantic space training phase. Projecting all the test image in bulk took under 2 minutes, and it took about 5 minutes to generate annotations for all the 13000 images; so, in general, it took less than .05s to get from a list of visual terms to the suggested annotations for a single image.

Implementation. The semantic-space software is written in C and makes use of Doug Rohde's SVDLIBC [1] for efficiently performing the large sparse SVD. The feature detector and descriptor software is written in C and C++. The image processing components were driven by a standard UNIX make file, which enabled easy parallelisation.

4 Conclusions

For this task we applied an older technique for automatically annotating images using a semantic space. The performance of the technique in terms of EER and AUC is fairly competitive, however, the technique does not mesh well with the hierarchical scoring measure proposed for this task. The semantic space technique is reasonably computationally efficient; the most time is spent processing

Feature	**Time**
Difference-of-Gaussian detection + SIFT extraction	≈ 1.8s/image
MSER detection	≈ 0.1s/image
SIFT extraction on MSER	≈ 2.7s/image
Colour-SIFT extraction on MSER	≈ 1.0s/image
Vector quantisation	<0.1s per set of extracted features
Estimated total	*≈ 5.9s/image*

Table 5: Approximate timings for feature extraction on a typical image from the training set.

[1] `http://tedlab.mit.edu/~dr/SVDLIBC/`

the images to extract features. In our experiments, the use of the hierarchy did not lead to any improvement in the annotation quality.

Acknowledgements

The authors wish to thank the European Union, which supported this work under the Seventh Framework project LivingKnowledge (IST-FP7-231126) and the LiveMemories project, graciously funded by the Autonomous Province of Trento (Italy).

References

[1] Gertjan J. Burghouts and Jan-Mark Geusebroek. Performance evaluation of local colour invariants. *Computer Vision and Image Understanding*, 113(1):48 – 62, 2009.

[2] Jonathan S. Hare, Sina Samangooei, Paul H. Lewis, and Mark S. Nixon. Semantic spaces revisited: investigating the performance of auto-annotation and semantic retrieval using semantic spaces. In *ACM CIVR '08*, pages 359–368. ACM, July 2008.

[3] Jonathon S. Hare, Paul H. Lewis, Peter G. B. Enser, and Christine J. Sandom. A Linear-Algebraic Technique with an Application in Semantic Image Retrieval. In Hari Sundaram, Milind Naphade, John R. Smith, and Yong Rui, editors, *CIVR 2006*, volume 4071 of *LNCS*, pages 31–40. Springer, 2006.

[4] Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM.

[5] T K Landauer and M L Littman. Fully automatic cross-language document retrieval using latent semantic indexing. In *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pages 31–38, UW Centre for the New OED and Text Research, Waterloo, Ontario, Canada, October 1990.

[6] David Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, January 2004.

[7] Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In Paul L. Rosin and A. David Marshall, editors, *BMVC*. British Machine Vision Association, 2002.

[8] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *In CVPR*, pages 2161–2168, 2006.

[9] Stefanie Nowak and Peter Dunker. Overview of the CLEF 2009 Large Scale - Visual Concept Detection and Annotation Task. In *CLEF working notes 2009*, Corfu, Greece, 2009.

[10] Stefanie Nowak and Hanna Lukashevich. Multilabel classification evaluation using ontology information. In Claudia d'Amato, Nicola Fanizzi, Marko Grobelnik, Agnieszka Lawrynowicz, and Vojtech Svátek, editors, *Proceedings of the First ESWC Workshop on Inductive Reasoning and Machine Learning on the Semantic Web*, Heraklion, Greece, June 2009.

[11] J Sivic and A Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, October 2003.