SZTAKI @ ImageCLEF 2009

Bálint Daróczy István Petrás András A. Benczúr Zsolt Fekete
Dávid Nemeskey Dávid Siklósi Zsuzsa Weiner
Data Mining and Web search Research Group, Informatics Laboratory
Computer and Automation Research Institute of the Hungarian Academy of Sciences
{benczur, daroczyb, zsfekete, ndavid, petras, sdavid, weiner}@ilab.sztaki.hu

Abstract

Our approach to the ImageCLEF 2009 tasks is based on image segmentation, SIFT keypoints and Okapi BM25 based text retrieval. We use feature vectors to describe the visual content of an image segment, a keypoint or the entire image. The features include color histograms, a shape descriptor as well as a 2D Fourier transform of a segment and an orientation histogram of detected keypoints. We trained a Gaussian Mixture Model (GMM) to cluster the feature vectors extracted from the image segments and keypoints independently. The normalized Fisher gradient vector computed from GMM of SIFT descriptors is a well known technique to represent an image with only one vector. Novel to our method is the combination of Fisher vectors for keypoints with those of the image segments to improve classification accuracy. We introduced training and correlation based combining methods to further improve classification quality.

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 segmentation, SIFT, Gaussian mixtures, Okapi BM25, rank aggregation

1 Introduction

In this paper we describe our approach to the ImageCLEF Photo, WikiMediaMM and Photo Annotation 2009 evaluation campaigns [11, 17, 12]. ImageCLEF Photo is over 498,920 images from Belga News Agency, WikiMediaMM is over the INEX MM image database of approximately 150,000 images and Photo Annotation is over the MIR Flickr 25.000 image dataset. The images are associated with unstructured and noisy textual annotations in English. The first two campaigns are ad-hoc image retrieval tasks: find as many relevant images as possible from the image collections. The third campaign requires image classification into 53 concepts organized in a small ontology.

The key feature of our solution in both cases is to combine text based and content based image retrieval. Our method is similar to the method we applied last year for ImageCLEF Photo [6]. Our CBIR method is based on segmentation of the image and on the comparison of features of

the segments. We use the Hungarian Academy of Sciences search engine [2] as our information retrieval system that is based on Okapi BM25 [16] and query expansion by thesaurus.

2 Image processing

We transform images into a feature space both in order to define their similarity for ad hoc retrieval and to apply classifiers over them for annotation. For image processing we deploy both SIFT keypoints [8] and image segmentation [5, 14, 4, 9]. While SIFT is a standard procedure, we describe our home developed segmenter in more detail below.

2.1 Segmentation

Our segmentation algorithm is based on a graph of the image pixels where the eight neighbors of a pixel are connected by edges. The weight of an edge is equal to the Euclidean distance of the pixels in the RGB space. We proceed in the order of increasing edge weight as in a minimum spanning tree algorithm except that we do not merge segments if their size and the similarity of their boundary edges are above a threshold. In the algorithm we use the notation

$$B(S_1, S_2) = \text{average weight of edges connecting } S_1 \text{ and } S_2.$$

The algorithm consists of several iterations of the above minimum spanning tree type procedure. In the first iteration we join sturdily coherent pixels into segments. In further iterations we gradually increase the limits in order to enlarge segments and reach a required number of them.

The algorithm is called with three parameters τ_1, τ_2 and τ_3 where the first is initialized to be the difference of the minimal and maximal edge weight in the graph while the other two are chosen to have values 40 and 50, respectively.

Algorithm 1 Algorithm Segmentation($I_{\text{src}}, \tau_1, \tau_2, \tau_3$).

for all pixels p **do**
  define segment $S_p = \{p\}$
  $\tau(S_p) \leftarrow \tau_1$

{Joining sturdily coherent pixels}
for all neighboring pixel pairs (p, q) in the order of edge weight **do**
  if $S_p \neq S_q$ and $\min\{\tau(S_p), \tau(S_q)\} > B(S_p, S_q)$ **then**
    $S_p \leftarrow S_p \cup S_q$
    $\tau(S_p) \leftarrow \frac{\tau(S_p) * |S_p| + \tau(S_q) * |S_q|}{|S_p| + |S_q|} + B(S_p, S_q)$

{Segment enlargement}
while we reach the prescribed number of segments **do**
  for all neighboring pixel pairs (p, q) in the order of edge weight **do**
    if $S_p \neq S_q$ and $min(|S_p|, |S_q|) < \tau_2$ and $B(S_p, S_q) < \tau_3$ **then**
      $S_p \leftarrow S_p \cup S_q$
  $\tau_2 \leftarrow \tau_2 * 1.2$ and $\tau_3 = \tau_3 * 1.3$

2.2 Feature extraction

We performed colour, shape, orientation and texture feature extraction over the segments and environment of keypoints of images. This resulted in approximately $0.5 - 7$ thousand keypoint descriptors in 128 dimensions and in approximately 0.2 thousand segment descriptors in 350 dimensions. The following features were extracted for each segment: mean RGB histogram; mean HSV histogram; normalized RGB histogram; normalized HSV histogram; normalized contrast histogram; shape moments (up to 3^{rd} order); DFT phase and amplitude.

Table 1: WikiMediaMM ad hoc search evaluation.

	MAP	P10	P20
Image+Text	0.1699	0.2867	0.2389
Text	0.1676	0.2911	0.2411
Image+Text+Thesaurus	0.1604	0.2778	0.2200
Text+Thesaurus	0.1583	0.2667	0.2122
Image	0.0068	0.0244	0.0144

2.3 Image Similarity

For ad hoc image retrieval we considered segmentation based image similarity only. We extracted features for color histogram, shape and texture information for every segment. In addition we used contrast and 2D Fourier coefficients. The discrete Fourier transformation was sampled along a zig-zag order, i.e. the low frequency components were included. An asymmetric distance function is defined in the above feature space as

$$d(D_i, D_j) = \sum_k \min_\ell \operatorname{dist}(S_{ik}, S_{j\ell})$$

where $\{S_{dt} : t \geq 1\}$ denotes the set of segments of image D_d. Finally image similarity rank was obtained by substracting the above distance from a sufficiently large constant.

3 The base text search engine

We use the Hungarian Academy of Sciences search engine [2] as our information retrieval system based on Okapi BM25 ranking [16] with the proximity of query terms taken into account [15, 3]. We deployed stopword removal and stemming by the Porter stemmer. We extended of stop word list with terms such as "photo" or "image" that are frequently used in annotations but does not have a distinctive meaning in this task.

We applied query term weighting to distinguish definite and rough query terms, the latter may be obtained from the topic description or a thesaurus. We multiplied the BM25 score of each query term by its weight; the sum of the scores gave the final rank.

We used a linear combination of the text based and image similarity based scores for ad hoc retrieval. We considered the text based score more accurate used small weight for the content based score.

4 The WikipediaMM Task

We preprocessed the annotation text by regular expressions to remove author and copyright information. We made no differentiation between the title and the body of the annotation.

Since file names often contain relevant keywords and also often as substring, we gave score proportional to the length of the matching substring. Since the indexing of all substrings is infeasible, we only performed this step for those documents that already matched at least one query term in their body.

For the WikipediaMM task we also deployed query expansion by an online thesaurus[1]. We added groups of synonyms with reduces weight so that only the score of the first few best performing synonym was added to the final score to avoid overscoring long lists of synonyms.

As seen in Table 1, our CBIR score improved performance in terms of MAP for the price of worse early precision. In this experiment expansion by thesaurus did not help.

[1] `http://thesaurus.com/`

Table 2: ImageCLEF Photo ad hoc search evaluation.

	F-measure	P5	P20	CR5	CR20	MAP
Text CT	0.6449	0.5	0.64	0.5106	0.6363	0.49
Text	0.6394	0.52	0.68	0.4719	0.6430	0.50
Image+Text CT	0.6315	0.49	0.64	0.4319	0.6407	0.48
Image	0.1727	0.02	0.03	0.2282	0.2826	0

5 The Photo Retrieval Task: Optimizing for Diversity

We preprocessed the annotation text by regular expressions to remove photographer and agency information. This step was in particular important to get rid of the false positives for Belgium-related queries as the majority of the images has the Belga News Agency as annotated source. Since the annotation was very noisy, we could only approximately cleanse the corpus.

As the main difference from the WikimediaMM task, since almost all queries were related to names of people or places, we did not deploy the thesaurus. Some of the topics had description (denoted by CT in the topic set as well as in Table 2) that we added with weight 0.1.

We modified our method to achieve greater diversity within the top 20. For each topic in the ImageCLEF Photo set, relevant images were manually clustered into sub-topics. Evaluation was based on two measures: precision at 20 and cluster recall at rank 20, the percentage of different clusters represented in the top 20.

The topics of this task were of two different types and we processed them separately in order to optimize for cluster recall. The first set of topics included subtopics; we merged the hit lists of the subtopics by one by one. The last subtopic typically contained terms from other subtopics negated; we fed the query with negation into the retrieval engine.

The other class of topics had no subtopics; here we proceeded as follows. Let Orig(i) be the ith document ($0 \leq i < 999$) and OrigSc(i) be the score of this element on the original list for a given query Q_j. We modified these scores by giving penalties to the scores of the documents based on their Kullback-Leibler distance. We used the following algorithm.

Algorithm 2 Algorithm Re-ranking

1. New (0) = Orig(0) and NewSc(0) = OrigSc(0)
2. For $i = 1$ to 20
   (a) New(i) = $\text{argmax}_k\{\text{CL}_i(k) \mid i <= k < 999\}$
   (b) NewSc(i) = $\max\{\text{CL}_i(k) \mid i <= k < 999\}$
   (c) For $\ell = 0$ to $(i-1)$
   NewSc(ℓ) = NewSc(ℓ) + $c(i)$

Here $\text{CL}_i(k) = \text{OrigSc}(k) + \alpha \sum_{l=0}^{i-1} \text{KL}(i,k)$, where α is a tunable parameter and KL(i,k) is the Kullback-Leibler distance of the ith and kth documents. We used a correction term $c(i)$ at Step (2c) to ensure that the new scores will be also in descending order.

6 The Photo Annotation Task

The Photo Annotation data consisted of 5000 annotated training and 13000 test images. Our overall procedure is shown in Fig. 1. We used the bag-of-visual words (BOV) generative approach in combination with the Fisher kernels method for images [13, 1]. As a first step we extracted low level features from each image. These features include the SIFT key points and the color image segment descriptors such as shape, color histogram as described in Section 2.2.

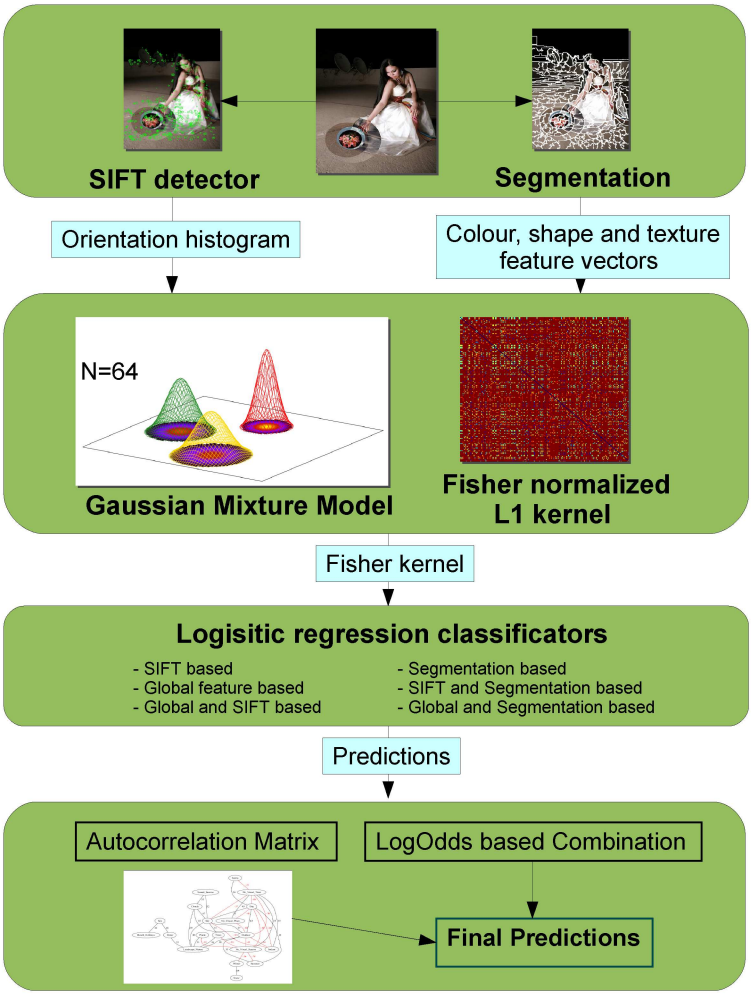

Figure 1: Our image annotation procedure

We produced a global visual vocabulary that approximate the per-image distribution of the low level features by clustering with a 64 dimensional GMM. First we obtained a variable number of visual words per image that we processed by Fisher kernels. The resulting kernel from different feature combinations were used as training input for a binary linear classifier (L_2 logistic regression). We used a held-out set to rank each row from the Fisher kernel. After computing the results for all of the 53 concepts, a matrix of dimensionality $N \times 53$ holds the concept detection results, where N is the number of images.

The concept detection results from different kernels can be combined. We followed two approaches. The first one described in Section 6.2 exploits the connection between the concepts of the training annotation while the second one (Section 6.3) applies another round of training to learn the best combination of the individual concept detectors.

6.1 Feature generation and modeling

To reduce the size of the feature vectors we modeled them with 64 Gaussians. The classical EM algorithm with diagonal covariance matrix assumption was used for the computation of the mixture parameters. To get fixed sized image descriptors we computed $g-1+g\times D\times 2$ dimensional normalized Fisher vectors per images [13, 1], where $D = 128$ is the dimension of the low level feature vectors. The $t \times t$ Fisher kernel matrix contained the L1 distances of all training images

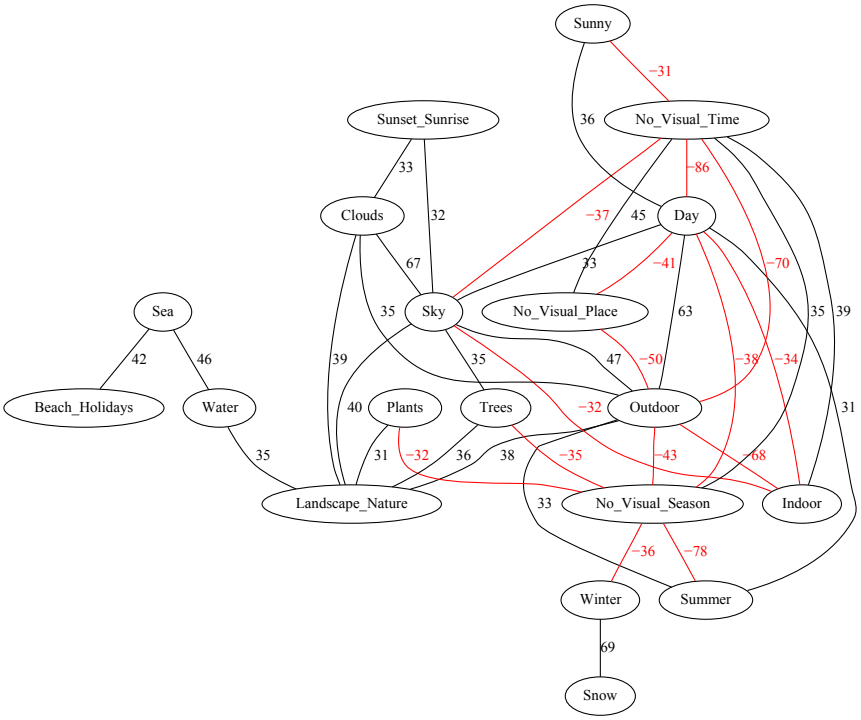

Figure 2: The auto-correlation matrix of the training annotation visualized as a graph. This graph was used to re-weight the output of the predictors. Positive weights mean positive correlation between concepts. Connection can be expressed in verbal form, e.g. "*Landscape_Nature* correlates with *Trees, Sky, Water, Plants, Clouds* as well as with *Outdoor*"

from themselves. There are $t = 5000$ training images. We computed the Fisher kernels for several low level feature type combinations. Such combinations were: SIFT+image segments, SIFT+global image features, etc. We used the resulting Fisher kernels for training binary linear classifiers (L2-regularized logistic regression classifier from the LibLinear package [7]) for each of the $k = 53$ concepts. For prediction we used the $s \times t$ kernel matrix with the trained linear classifiers, where $s = 13000$ denotes the number of test images.

6.2 Correlation based combination

From the annotations of the training images we computed the auto-correlation matrix (Fig. 2). Using this matrix we exploited the common knowledge of annotations about the relationship between the concepts. With this matrix we reweighted the output of the predictors. Let us denote A the $t \times k$ annotation matrix. Each entry of A is either 0 or 1. Moreover, let $C = [c_{ij}]$ be the $k \times k$ symmetric correlation matrix where $c_{ij} = corr\left(a_i, a_j\right)$, a_i is the i^{th} column of A, $corr\left(x, y\right) = cov\left(x, y\right) / \left(std\left(x\right) \cdot std\left(y\right)\right)$ is the normalized correlation coefficient. Let P denote the $t \times k$ matrix composed from the outputs of the predictors. Rows correspond to images, while columns correspond to concepts. The combined prediction is computed

$$P_C = PC$$

The improvement is shown in Table 3.

6.3 Log-odds based combination

Our combination of the classifiers is inspired by the log-odds averaging by Lynam and Cormack [10]. We first made a 10-fold crossvalidation on the training data to score every image by every

Nr.	Concept	AUC	AUC comb.	LogOdds c.	Nr.	Concept	AUC	AUC comb.	LogOdds c.
0	Partylife	0.73957	0.72941	0.68350	27	Day	0.74645	0.76309	0.76555
1	Familiy_Friends	0.73815	0.73811	0.71786	28	Night	0.86187	0.86187	0.87458
2	Beach_Holidays	0.81424	0.85588	0.79171	29	No_Visual_Time	0.74712	0.75651	0.74943
3	Building_Sights	0.80933	0.80933	0.79170	30	Sunny	0.72356	0.73187	0.72782
4	Snow	0.80294	0.75331	0.79255	31	Sunset_Sunrise	0.92588	0.92588	0.92026
5	Citylife	0.78624	0.78624	0.75672	32	Canvas	0.73779	0.73779	0.73235
6	Landscape_Nature	0.89293	0.89293	0.88607	33	Still_Life	0.74807	0.75257	0.72224
7	Sports	0.59259	0.59820	0.51282	34	Macro	0.70740	0.70740	0.69002
8	Desert	0.83242	0.83242	0.74518	35	Portrait	0.77825	0.77825	0.73150
9	Spring	0.75002	0.75002	0.68572	36	Overexposed	0.76191	0.76191	0.72049
10	Summer	0.77211	0.77541	0.73943	37	Underexposed	0.86044	0.86044	0.88017
11	Autumn	0.77894	0.77894	0.73783	38	Neutral_Illumination	0.79114	0.79114	0.79946
12	Winter	0.79843	0.79843	0.80648	39	Motion_Blur	0.66064	0.66064	0.62978
13	No_Visual_Season	0.76659	0.75141	0.69890	40	Out_of_focus	0.79453	0.74420	0.72747
14	Indoor	0.72815	0.69309	0.71512	41	Partly_Blurred	0.82296	0.82296	0.78099
15	Outdoor	0.81166	0.80824	0.80551	42	No_Blur	0.79566	0.79566	0.73280
16	No_Visual_Place	0.72283	0.72283	0.64580	43	Single_Person	0.71757	0.71757	0.62987
17	Plants	0.78703	0.79857	0.78064	44	Small_Group	0.66305	0.66305	0.66832
18	Flowers	0.79807	0.81733	0.80806	45	Big_Group	0.78595	0.78595	0.72330
19	Trees	0.82024	0.83775	0.85227	46	No_Persons	0.72840	0.72840	0.72431
20	Sky	0.87776	0.87776	0.85901	47	Animals	0.74320	0.74320	0.70871
21	Clouds	0.88680	0.89788	0.88929	48	Food	0.83345	0.84414	0.83605
22	Water	0.80937	0.80937	0.79304	49	Vehicle	0.72292	0.73282	0.70768
23	Lake	0.79240	0.85950	0.73959	50	Aesthetic_Impression	0.63817	0.63817	0.62796
24	River	0.74148	0.74148	0.70190	51	Overall_Quality	0.63330	0.63330	0.55105
25	Sea	0.87256	0.89580	0.87399	52	Fancy	0.56859	0.56859	0.52613
26	Mountains	0.85911	0.85911	0.85493					
	mean AUC	**0.7713**	**0.7731**	**0.7463**		**increase:**		**0.0018**	**-3.3407**

Table 3: Results of the predictors on the test data without and with combinations. Column AUC contains the output of the predictors using SIFT and segmentation feature vectors. Next column shows the results after combining the previous column with the autocorrelation matrix of the training annotation data. "log-odds" column contains the output of the combination of predictors using log-odds. The following prediction methods were combined: segmentation and SIFT, global features only, SIFT only, segmentation only , global and SIFT features, global features and segmentation

classifier. Then for every classifier we calculated the log-odds as a feature by taking the logarithm of the fraction of the number of positive images with lower score over the number of negative images with higher score. Finally, we trained a logit-boost classifier over this feature set. The predictors were trained with the following feature sets: segmentation and SIFT (two fine tuned runs); global features only; SIFT only; segmentation only; global and SIFT features; global features and segmentation.

7 Conclusion

- For image classification, we successfully combined a pure keypoint based and a region based method, two image processing algorithms that complement each other. Further improvement could be to include the hierarchical relationship of the concepts into the combination procedure that would result in a directed graph to describe $Concept_A \rightarrow Concept_B$ relation-

Table 4: ImageCLEF2009-PhotoAnnotation results

	EER	AUC
Segmentation	0.346106	0.707860
SIFT	0.322632	0.733264
SIFT + Segmentation	0.296315	0.771324
SIFT + Segmentation + Cross	0.291718	0.773133
Log Odds combination	0.304113	0.746300

ships.

- For image retrieval our content based score improved the text score in combination. The use of the thesaurus and other query expansion techniques needs further analysis and refinement.
- We took minimal effort for optimizing for diversity; while our results were strong in MAP, optimization with stronger parameters could have helped.

References

[1] J. Ah-Pine, C. Cifarelli, S. Clinchant, G. Csurka, and J.M. Renders. XRCE's Participation to ImageCLEF 2008. In *Working Notes of the 2008 CLEF Workshop*, 2008.

[2] András A. Benczúr, Károly Csalogány, Eszter Friedman, Dániel Fogaras, Tamás Sarlós, Máté Uher, and Eszter Windhager. Searching a small national domain—preliminary report. In *Proceedings of the 12th World Wide Web Conference (WWW)*, 2003.

[3] Stefan Büttcher, Charles L. A. Clarke, and Brad Lushman. Term proximity scoring for ad-hoc retrieval on very large text collections. In *SIGIR '06*, pages 621–622, New York, NY, USA, 2006. ACM Press.

[4] Chad Carson, Serge Belongie, Hayit Greenspan, and Jitendra Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(8):1026–1038, 2002.

[5] Yixin Chen and James Z. Wang. Image categorization by learning and reasoning with regions. *J. Mach. Learn. Res.*, 5:913–939, 2004.

[6] Bálint Daróczy, Zsolt Fekete, Mátyás Brendel, Simon Rácz, András Benczúr, Dávid Siklósi, and Attila Pereszlényi. Cross-modal image retrieval with parameter tuning. In Carol Peters, Danilo Giampiccol, Nicola Ferro, Vivien Petras, Julio Gonzalo, Anselmo Peñas, Thomas Deselaers, Thomas Mandl, Gareth Jones, and Nikko Kurimo, editors, *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, Lecture Notes in Computer Science, Aarhus, Denmark, September 2008 (printed in 2009).

[7] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. LIBLINEAR: A library for large linear classication. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[8] D.G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, volume 2, pages 1150–1157. Corfu, Greece, 1999.

[9] Qin Lv, Moses Charikar, and Kai Li. Image similarity search with compact data structures. In *CIKM '04: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, pages 208–217, New York, NY, USA, 2004. ACM Press.

[10] T.R. Lynam, G.V. Cormack, and D.R. Cheriton. On-line spam filter fusion. *Proc. of the 29th international ACM SIGIR conference on Research and development in information retrieval*, pages 123–130, 2006.

[11] Stefanie Nowak and Peter Dunker. Overview of the CLEF 2009 large scale visual concept detection and annotation task. In *Working Notes for the CLEF 2009 Workshop*, 2009.

[12] M Paramita, M Sanderson, and P Clough. Diversity in photo retrieval: overview of the ImageCLEFPhoto task 2009. In *Working Notes for the CLEF 2009 Workshop*, 2009.

[13] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*, pages 1–8, 2007.

[14] B. G. Prasad, K. K. Biswas, and S. K. Gupta. Region-based image retrieval using integrated color, shape, and location index. *Comput. Vis. Image Underst.*, 94(1-3):193–233, 2004.

[15] Yves Rasolofo and Jacques Savoy. Term proximity scoring for keyword-based retrieval systems. In *ECIR*, pages 207–218, 2003.

[16] Stephen E. Robertson and Karen Sparck Jones. Relevance weighting of search terms. In *Document retrieval systems*, pages 143–160. Taylor Graham Publishing, London, UK, UK, 1988.

[17] Theodora Tsikrika and Jana Kludas. Overview of the WikipediaMM task at ImageCLEF 2009. In *Working Notes for the CLEF 2009 Workshop*, 2009.