Multimodal Photo Retrieval through Finding Similar Documents Enhanced with Visual Clues – a Baseline Method

Bartosz Broda, Mariusz Paradowski, Halina Kwanicka
Institute of Informatics, Wroclaw University of Technology
27 Wybrzee Wyspiaskiego 50-370 Wrocaw, Poland
{bartosz.broda, mariusz.paradowski, halina.kwasnicka}@pwr.wroc.pl

Abstract

Image retrieval till today is one of the most challenging problems in computer science. Even though there are lots of researches performed around the World, an efficient, user friendly image retrieval system still seems to be an unachievable goal. *ImageCLEF Photo Retrieval Track* allows to compare various approaches to this challenging problem. In this paper we present a staring point of our research, connected to a joint Polish–Singaporean research project, titled: *Framework for Visual Information Retrieval and Building Content-based Visual Search Engines*. Various techniques, published in the literature are gathered and orchestrated together. A reference image retrieval system is build, supporting both image queries, text queries and joint text-image queries. In our work we have tried to capture state-of-the-art in text and image retrieval.

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 Management**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Information retrieval, Image Retrieval, Integrated Region Matching, Mallows Distance, Document similarity, Vector Space Model

1 Introduction

In this paper we present work performed during our participation in ImageCLEF Photo Retrieval Task. The presented approach is actually a baseline, the starting point of our research. The work is done as a part of joint Polish–Singaporean research project, titled *Framework for Visual Information Retrieval and Building Content-based Visual Search Engines*. The aim of our work is to build image retrieval methods, incorporating both image matching techniques, 'soft' image distance measures and knowledge based methods. As this is a starting point, all presented techniques are not a novelty. They are already presented in the research literature for at least several

years. However, the key difficulty is such combination of these known approaches to have a viable and efficient image retrieval system.

We participate in *ImageCLEF* contest for the first time. Our goal is actually to check: *Where we are?*, comparing to other, often much larger and more experienced research teams. We have achieved the goal by preparing a simple image retrieval system, using state-of-the-art techniques described in the literature. On the other hand, since the contest took place new ideas have appeared and are currently under heavy research. We hope that these ideas will be verified by the next contest – *ImageCLEF 2010*.

2 Task Description

As the name suggest *ImageCLEF 2009 Photo Retrieval Task*[1] (henceforth, ICPRT2009) is one of the tasks of information retrieval. The goal of information retrieval, and more specific *image retrieval*, is to satisfy users information needs — finding images that are relevant to user queries. In ICPRT2009 those needs are specified by detailed description of user queries in form of topics. Topics were divided into two groups containing 25 queries. In first group every topic is described by title and number of clusters. A cluster has title, description and image example. The second group of topics had only title and several image examples. Figure 1 shows one of the topics developed for ICPRT2009 competition.

(a) Cluster title: kim gevaert. Description: Relevant images will show photographs of Kim Gevaert. Images showing her and other people are relevant if she is shown in the foreground. Other images where Kim is in the foreground are irrelevant.

(b) Cluster title: agfa gevaert. Description: Images are relevant if they contain photographs of the Agfa-Gevaert company. Relevant images include those showing the logos, buildings or any aspects of the company.

(c) Cluster title: hellebaut gevaert. Description: Relevant images show photographs of Hellebaut and Kim Gevaert. Images showing only one of them are not relevant.

Figure 1: Example of user query in form of a topic. The topic has number 7, is titled *gevaert* and contains only three clusters.

One of the main goals of the ICPRT2009 competition is to encourage participants to focus on diversity in retrieved collection of images. User queries often do not contain the true intentions, delivered in an machine understandable (or even human understandable) form. They are ambiguous by definition. Taking this into account, diversity is one of the approaches to diminish the problem of query ambiguity. In classical information retrieval, when users searches for word "bank", it is not possible to determine whether user wants to find information about river banks

[1] `http://imageclef.org/2009/photo`

or financial institutions[2]. In image retrieval this corresponds to searching for a Formula 1 bolid by presenting image of a bolid during the race (probably accompanied by short description). Does the user want to find any images of bolids or should the pictures be taken during the race? Perhaps user is interested in pictures of only one team or only from the same event? One of the approaches to this problem is to focus on finding all potentially relevant images that encompasses as broad area of topics as possible.

One of the most interesting things about the ICPRT2009 is that the work is performed on the real image database from Belga News Agency and the topics created on the basis of analysis of Belga query logs.

The image database consists of 498,920 images. Every image is accompanied by caption, or annotation, i.e., a few English sentences describing image content. The format of captions is not formally standardized, so basically a caption can contain anything. Fortunately, one can observe a pattern that is usually followed: at the beginning the uppercase letters contains image identifier, date and place followed by description of image content (normal case). Caption is usually ended by attribution of authorship, again written in upper case. Figure 2 shows example of images with caption.

The size of the database introduces interesting efficiency problems to solve, especially with regard to image processing. At first, image feature vectors need to be computed from the image database. Afterward, image distance calculations need to be made. As it is presented later on in the paper, a single distance calculation requires solving an optimization problem. Such process, repeated for hundreds of thousands of images, is very time consuming. On the other hand, processing almost 500,000 documents (captions) might seems difficult. On the contrary, due to short length of captions, it is rather easy and straightforward process. It is worth noting that file with captions in raw text format has only 162MB, which is little comparing to other information retrieval tasks that has to deal with hundreds of gigabytes of data, e.g., [2, 16]. Creating all the necessary data structures in our case takes from 7 to 15 minutes on commodity PC depending on the experiment setting.

3 Development of the Baseline Method

It is very difficult to define visual similarity between images in a formal way. This is because one would have to know semantics of a given image. Note that there can be a huge difference between images that are visually and semantically similar. Two images of silver cars can be visually very similar, even if one of the images shows an Audi and the other one shows only a toy car, a model of Mercedes. User searching for pictures of Audi would not be satisfied with images of toy cars. On the other hand if the user would be interesting in any silver car picture taken from the side of the car he might be perfectly satisfied with that result. We think that extraction of semantic clues from text is significantly easier then from images so for the ICPRT2009 we assumed model that uses textual features as a primary knowledge source. After initial processing of textual data, visual data is used to refine the results presented to user.

There is a plethora of methods for textual information retrieval [10]. Combining this with techniques using visual data in image retrieval [3] gives very large number of possible ways to approach ICPRT2009.

As mentioned earlier in the paper, the presented research is a part of joint Polish–Singaporean research project, titled: *Framework for Visual Information Retrieval and Building Content-based Visual Search Engines*. One of aims of our project is to develop methods for finding visually similar images to a given one using only visual information. We divided the work into two stages: creating ranked list of similar images for every image in a topic and combining those lists considering all the images in the topic. We separated textual processing from visual, because we wanted to focus on robust and reliable techniques from both visual and textual point of views separately. After that we have developed a method for combining both knowledge sources. Our main aim in this research is to create prototype method for image retrieval for which precision of retrieval is

[2] At least without asking the user for clarification or introduction of additional techniques like user profile.

(a) BRU199 - 20031012 - METTET, BELGIUM : Illustration picture shows a pilot making smoke during a burn-put wheelie at the Superbiker Grand Prix of Mettet, Sunday 12 October 2003, in Mettet. BELGA PHOTO JOHN THYS

(b) The early morning rising sun over the Bavarian town of Marktoberdorf gives a golden glow to the cloud cover of Tuesday 2nd December 2003.. This area just short of the Alpine region continues to enjoy extraordinary mild temperatures with 16degrees c. recorded yesterday. EPA/Karl-Josef Hildenbrand COLOR

(c) Japan's maglev train setting a world speed record on an experimental track in Yamanashi Province, Tuesday 02 Decemeber 2003. The three-car magnetically levitated train reached a maximum speed of 581 kilometers per hour with technicians on board, according to Central Japan Railway Co. (JR Tokai) and the government-affiliated Railway Technical Research Institute, who operates the experimental train. EPA/EVERETT KENNEDY BROWN

(d) LHT22 - 20010223 - LAHTI, FINLAND : From L to R Germany's Martin Schmitt, silver medal, Poland's Adam Malysz, gold medal and Austria's Martin Hoellwarth, bronze medal, jubilate on the podium after the K 90 ski jump final at the Nordic World Ski Championships in Lahti on Friday, 23 February 2001. EPA PHOTO EPA-ANJA NIEDRINGHAUS

Figure 2: Example of image with caption (annotation) from Belga News Agency collection.

the most important factor. As mentioned earlier, we are treating this work as a test ground for developing a baseline method upon which we are going to build more complete solutions later on.

3.1 Using Visual Clues

Similarity search in image domain always have and still is a great challenge [3]. Despite there are hundreds of different approaches proposed in the literature, there is no general one working for a large domain of images. To make an image retrieval system, one has to decide about three key components: *image distance function*, *image segmentation method* and *feature extraction method*. Proper selection of all these components, so they fit together, is a difficult problem.

3.1.1 Distance function

In our work we have examined several methods widely discussed in the literature. *Local image distances* are our major interest topic. Local image distance is a distance calculated between

individual segments of images and later on transformed into an image distance. Such distance operates on sets of feature vectors (they are not ordered), instead on single feature vectors. This means that a single image I has to be defined in terms of its segments (and feature vectors) $i_k : k = \{1, ..n\}$ as follows:

$$I = \{i_1, i_2, ..., i_n\}. \tag{1}$$

As a result of our research the decision is made to use a variant of *Mallows distance*, called *Integrated Region Matching* [8, 17]. The method has been proposed by Wang in 2001 as a part of *SIMPLIcity* image retrieval system. Our experiments have shown that till today it is one of the most effective means of image retrieval. The distance function is defined as an optimization problem:

$$D(I, J) = \min_{S=[s_{ij}]} \sum_{i \in I} \sum_{j \in J} s_{ij} d(i, j), \tag{2}$$

constrained by:

$$s_{ij} \geq 0, \quad 1 \leq i \leq |I|, \quad 1 \leq j \leq |J|, \quad \sum_{i=1}^{|I|} \sum_{j=1}^{|J|} s_{ij} = \sum_{i=1}^{|I|} p_i = \sum_{j=1}^{|J|} q_j = 1, \tag{3}$$

$$\sum_{j=1}^{|J|} s_{ij} = p_i, \quad 1 \leq i \leq |I|, \quad \sum_{i=1}^{|I|} s_{ij} = q_j, \quad 1 \leq j \leq |J|, \tag{4}$$

where:
I, J – feature vector sets,
i, j – single vectors belonging to I and J, respectively,
$S = [s_{ij}]$ – significance matrix, the search space of the optimization methods,
p_i – probability of a region $i \in I$ (usually equal to segment relative size),
q_j – probability of a region $j \in J$ (usually equal to segment relative size),
$d(i, j)$ – vector distance measure, usually Euclidean distance.

Of course the optimization problem itself is very challenging and finding a global solution is simply not feasible. *Integrated Region Matching* approach is in fact an iterative greedy algorithm of image segment pairing. Distance calculation between two images does not take too much time, however calculating it for the whole database requires much computational power.

3.1.2 Image segmentation

The second mentioned key component is image segmentation. Processed image database is a general type database, containing various kinds of images, encompassing many visual domains. For such databases supervised image segmentation approach would have to consider all these domains and is not a good choice. This means, that an unsupervised image segmentation, with all its advantages and disadvantages has to be selected.

Unsupervised methods may be even further divided into *block-based approaches* (fixed image cuts) and *region-based approaches* (segmentation in a classical sense). Block-based methods have been adopted to automatic image annotation as an effective means of image segmentation [4]. Our earlier research in *automatic image annotation* [6, 11] also confirmed this observation. Generated feature vectors are much less prone to changes due to slight changes of image content.

The chosen image segmentation approach is a regular grid method. The grid itself has 5×5 dimensions for every image in the database. This means that each image is split into 25 identical in size, rectangular blocks.

3.1.3 Feature extraction

The last key component are image features. In this research we have focused on three types of image segment features: *location*, *color* and *texture*. Such approach is popular among image retrieval and automatic image annotation research, e.g. [4]. Let us now describe features we are using. Location related features are:

- normalized region size (which is constant due to grid segmentation),
- region average x and y coordinates.

Color features are rather straightforward. Only two basic color models are used:

- region intensity means *red*, *green* and *blue* (RGB color model),
- region intensity standard deviations *red*, *green* and *blue*,
- region intensity means *hue*, *brightness* and *saturation* (HSV color model),
- region intensity standard deviations *hue*, *brightness* and *saturation*.

Texture features are much more complex. They include result of image processing by *Sobel edge detector*, Hessian-based edge detector[3] and concurrence matrices:

- region intensity of *Sobel edge detector* means for *red*, *green* and *blue* channels,
- region intensity of *Sobel edge detector* standard deviations for *red*, *green* and *blue* channels,
- region intensity of *Hessian-based edge detector* means for *red*, *green* and *blue* channels,
- region intensity of *Hessian-based edge detector* standard deviations for *red*, *green* and *blue* channels,
- region concurrence matrix values: correlation, entropy, homogeneity, contrast, dissimilarity, energy and sum of all values for *red*, *green* and *blue* channels.

3.2 Using Textual Data

The idea for textual retrieval is similar to described in previous section for visual features. First we extract textual features for each image in the collection and than we use similarity measure to compare pairs of images. As this is baseline method for our group, we assumed only very shallow processing without usage of elaborate language processing tools. Also we used classic representation of textual data, namely *Vector Space Model* (VSM) [10, 13].

After initial inspection of both the topics and images captions we decided to use only images caption. We assumed that captions describe the image in a better way than cluster descriptions for VSM. For humans, cluster description is more informative, but it would require deeper level of processing and usage of more elaborate language processing tools. E.g., we would need to resolve negations for searching for terms that are irrelevant[4], prepare special stop-list, etc.

We treat each caption as a document. After preprocessing the documents are represented in VSM. We use cosine as a measure of similarity. Preprocessing involves two steps. We use a stop-list, i.e., a list of a few hundred words that do not contribute much to semantics of a document, like prepositions and articles. In some of our experiments we also use classic Porter stemming algorithm [12]. This allows us to both reduce index size and treat different morphological variants of a word (e.g., plural forms of a word) as one object called stem.

[3] Method taken from Bio-medical Imaging Java library, see: http://bij.isi.uu.nl/

[4] Creating heuristics that would work for 50 topics provided by ICPRT2009 organizer would be feasible, but we assumed that description are not constrained in any way.

As VSM is commonly known technique in natural language processing community we will outline only main points for readers with no experience in this area. The most important concept in VSM is that a document $\vec{D_i}$ is represented as a vector in an n-dimensional feature space, where n is a number of different terms found during indexing:

$$\vec{D_i} =< tf_{i,1}, tf_{i,2}, \ldots, tf_{i,n} > , \tag{5}$$

where $t_{i,j}$ is number of occurrences of term j in document i. As some frequencies of occurrences can be accidental it is best to use some weighting scheme. For this work we also focused on using classic and robust method, i.e., $tf.idf$ weighting scheme [13]. Instead of using raw term frequencies we use weights, that are calculated for term t in document d in the following way:

$$tf.idf_{t,d} = tf_{t,d} \cdot \log \frac{N}{df_t}, \tag{6}$$

where N is number of documents (captions), and df_t is number of documents containing the term t. Representation of captions as a vectors enables usage of many similarity (or distance) measures known form literature. We used cosine as a similarity measure for our baseline method, as it is shown many times that it copes well with high dimensional data spaces, including documents represented in VSM [1, 9, 10].

3.3 Combining Text with Images

As precision is our top priority rather than diversity, we did not come up with very elaborate techniques for problem of merging visual and textual features into the final model. Both, the visual and textual part of our system produces ranked lists of images for every image that appeared in topics accompanied by a score. For generation of textual list we used a cosine measure, which is a similarity measure giving values from 0 to 1 (in case of positive vector values). On the other hand, Integrated Region Matching (IRM) is an unbounded distance function. To convert distance to similarity we subtract from one normalized value of IRM function. After the conversion we simply multiply values of both functions if cosine is lower then threshold $t = 0.8$. We introduced the threshold in order to preserve almost perfect matches from textual phase.

The multiplication of both similarity functions results in a single ranked list of images for every image in topics. As every topic contain a few images, only one question remains: how to combine different similarity lists into one list for topic. In this step we also use very simple methods. First method we considered is *naive* joining of sorted similarity lists into one big list sorted by similarities. Duplicated images are removed from this list and the top 20 images are presented to user (or for evaluation). Second method is more *balanced*: we calculate k as a simple division of 20 by a number of clusters in the topic. Then from every list we select k best images and combine them all into resulting list. In case of duplicates we draw more images from randomly selected list among the lists containing duplicates. Resulting list is also sorted by similarity values.

4 Experimental Results

For evaluation we submitted five runs: four using both textual and visual clues and one for textual data only. These five runs differ only in the method of combining similarity lists for individual images into final list for topics. The other difference is usage of stemmer. Table 1 summarizes result achieved by our system. The measures used for evaluation are precision, cluster recall and F1-measure at different cutoff levels. For brevity, we show only two cutoff levels: at 10 and 20 documents. The former is used by organizers to rank all participating system in ICPRT2009 and the latter is the number of results we submitted for each topic.

Results for runs presented in the Table 1 differ very little. Not surprisingly with higher cutoff value the precision is lowered and the cluster recall is higher. Usage of stemmer has little impact on this task. Balanced scheme of similarity list joining seems to be better, especially in terms of cluster recall. Nevertheless, the difference is small.

Stemmer	Type	Modality	CR@10	CR@20	P@10	P@20	F@10	F@20
no	balanced	TXTIMG	0.5991	0.7300	0.8	0.71	0.6837	0,7356
yes	balanced	TXTIMG	0.5946	0.7529	0.8	0.72	0.6815	0,7351
no	naive	TXTIMG	0.5856	0.6717	0.79	0.72	0.6741	0,7291
no	naive	TXT	0.5856	0.6717	0.79	0.72	0.6741	0,7291
yes	naive	TXTIMG	0.5811	0.6743	0.8	0.72	0.6718	0,7286

Table 1: Results of experiments sorted by F1-measure at with cut-off at 10 position.

We approached the development of our baseline system with the proper method for evaluation the results in mind. That is why we decided to create independently from organizers a system for evaluation of precision. We did not consider evaluation of cluster recall, because as mentioned earlier we focus on the quality of retrieved list of similarities, especially with regard to visual similarity. Figure 3 shows example screen from application developed for supporting of manual annotation of results. All the results are stored in database and automatically retrieved when needed, so the pair of images has to be annotated only once even if it repeatedly occurs in different experiments. To shorten time needed for evaluation we added functionality for selection of statistically significant random sample from whole result set [5].

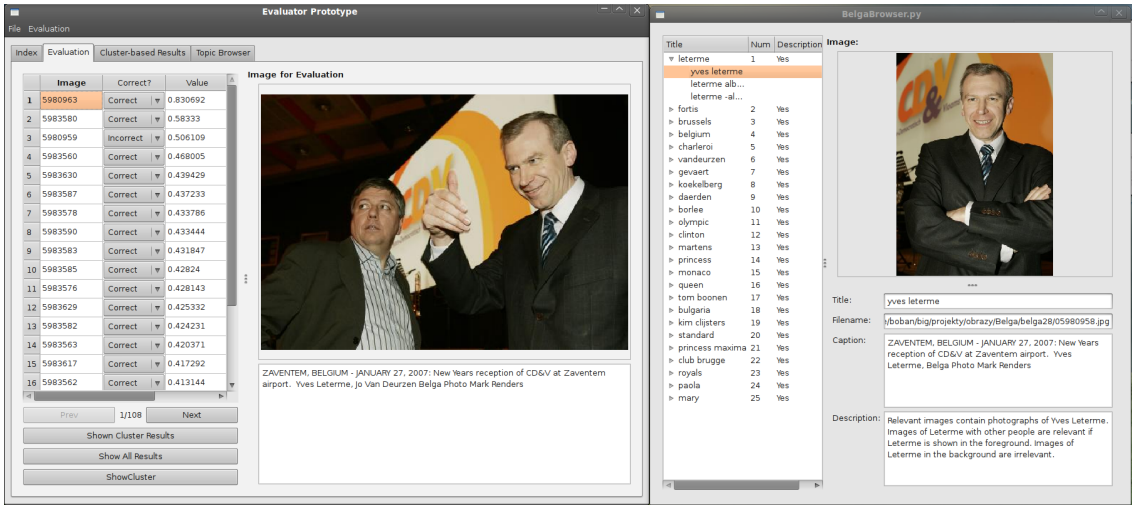


Figure 3: Example screen from evaluation application. On the left is the window showing 20 most similar ids of images to the image in cluster shown on the right returned by one of our methods. When user selects id of image for evaluation the image is shown in the frame. All the resulting images are already evaluated.

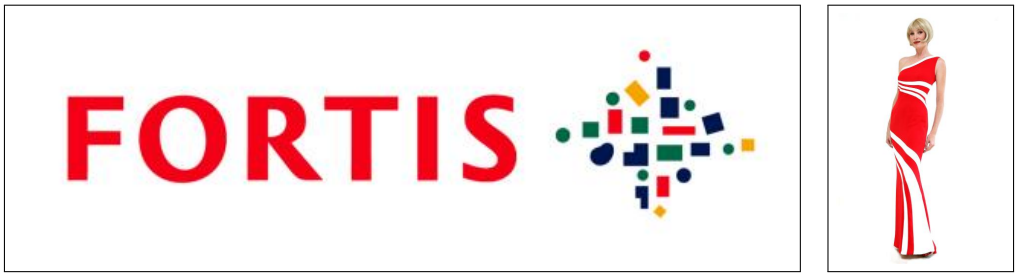
Surprisingly the result of evaluation that were obtained by our team were significantly lower then those prepared by ICPRT2009 organizers (see Tab. 1). Our best method did not achieved better precision than 0.56 (at cutoff of 20 documents). We think that main cause for this is that we evaluated n times more images for every topic then in submitted runs, where n is number of clusters in the topic. Another reason for lower precision is the problem that some images might have very vague caption or caption that did not corresponds exactly to the cluster description. For example, topics number 21 with the title "princess maxima" contains cluster with Princess Maxima appearing in different years. The cluster that describes year 2002 has an image of Princess Maxima during skiing. Naturally, textual part of the system is mislead into finding images containing other people skiing with similar captions. Also visually other people skiing are more similar than pictures from different time of the same year with Princess Maxima.

Figure 4 shows example of five best images retrieved for second cluster of topic 19 titled "justine

(a) Query image (b) 1^{st} result (c) 2^{nd} result

(d) 3^{rd} result (e) 4^{th} result (f) 5^{th} result

Figure 4: Example of our method: query image 4(a) with five most similar images.



(a) Query image

(b) Most similar image to query using IRM.

Figure 5: Example of visual similarity for one of the images in the topics.

henin kim clijsters". The cluster description says that relevant cluster should contain photographs of both Justine Henin and Kim Clijsters in the foreground to be relevant. This example shows an interesting case, because retrieved images are both semantically and visually similar to user query. On the other hand, we want to show the problems of usage of visual similarity based on high level statistical features derived from images segmented with grid for semantically oriented

user queries. Figure 5 shows query image containing Fortis Bank logo, and most similar image from whole Belga repository using only visual features. Without consideration for semantics of the image those two images are highly similar in our opinion. Problem shown on Fig. 5 is not an isolated case. This is the reason that we did not consider using only visual clues for submitting runs for ICPRT2009.

5 Conclusions and Further Work

We presented a system implementing baseline method for image retrieval that took part in ImageCLEF 2009 Photo Retrieval task. The method consists of three fairly independent components: visual, textual and merged. Splitting our system onto those parts has many benefits. We can improve each individual component independently, without affecting others, we are able to measure how much textual and visual modules contribute to the final results, etc.

In visual image retrieval part we have orchestrated three important components: image distance function, an image segmentation method and feature extraction approach. All these components play an important role is the system and are responsible for the quality of achieved visual processing.

Textual image retrieval by caption is based on the well-known Vector Space Model using $tf.idf$ weighting scheme and cosine as a measure of similarity between images captions [13]. We used also very naive technique for merging results of visual and textual components. Surprisingly we obtained rather high results with F1-measure lower only by 0.13 from the best system participating in this edition of Photo Retrieval Task. More surprisingly, when we consider only precision, which we were aiming at, the difference of our best approach from the most precise system in the competition is even lower: 0.05.

Obviously as this is our baseline system, there are many areas of improvement. From textual point of view one can try plethora of methods that were developed recently. Usage of deeper syntactic and semantic analysis can improve performance. Also, merging step in our method needs dramatical improvement. We consider using clustering algorithms for better partitioning both the data and the subset of the data containing only retrieved images. Developing a measure of certainty of retrieval for both visual and textual parts will lead to more intelligent ways of joining results using different modality.

As mentioned through the paper, we are intensely developing various approaches in image retrieval and automatic image annotation. It is worth mentioning the improved version of *MAGMA* [15] image annotation system, which also took part in ImageCLEF 2009 contest. We have also proposed a theoretical model of optimal automatic image annotation and its practical realization, called *Greedy Resulted Words Count Optimizer* [7,11]. Another concept being currently researched (but not yet published) is an extension of *Integrated Region Matching* matching. All presented image retrieval methods are based purely on spatial, color and texture features. We are also working on integration of those features with local features, such like *SIFT* or *Pattern-based approximations of Patches using Hough Transform* [14]. We hope to use at least some of the mentioned approaches in coming *ImageCLEF 2010*.

Acknowledgment This work is financed from the Ministry of Science and Higher Education Republic of Poland resources in 2008–2010 years as a Poland–Singapore joint research project 65/N-SINGAPORE/2007/0. It is supported by the DCS-Lab, which is operated by the Department of Distributed Computer Systems (DDCS) at the Institute of Informatics, Wroclaw University of Technology, Wroclaw, Poland.

References

[1] Bartosz Broda and Maciej Piasecki. Experiments in documents clustering for the automatic acquisition of lexical semantic networks for Polish. In Mieczysaw A. Kopotek, Adam Przepirkowski, Sawomir T. Wierzcho, and Krzysztof Trojanowski, editors, *Proceedings of the Sixteenth International Conference on Intelligent Information Systems*, Advances in Soft Computing, pages 203–212, Warsaw, 2008. Academic Publishing House EXIT.

[2] C. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2004 terabyte track. In *Proceedings of the 13th Text REtrieval Conference, Gaithersburg, USA*, 2004.

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

[4] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–1002–II–1009, 2004.

[5] Glenn D. Israel. Determining sample size. Technical report, University of Florida, 1992.

[6] Halina Kwasnicka and Mariusz Paradowski. Fast image auto-annotation with discretized feature distance measures. *Machine Graphics and Vision International Journal*, 15(2):123–140, 2006.

[7] Halina Kwasnicka and Mariusz Paradowski. Resulted word counts optimization – a new approach for better automatic image annotation. *Pattern Recognition*, 41(12):3562–3571, 2008.

[8] Jia Li, James Z. Wang, and Gio Wiederhold. Irm: Integrated region matching for image retrieval. pages 147–156, 2000.

[9] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 2001.

[10] C.D. Manning, P. Raghavan, and H. Schtze. *Introduction to information retrieval*. Cambridge University Press New York, NY, USA, 2008.

[11] Mariusz Paradowski. *Automatic Image Annotation as an Effective Method for Image Captioning (in Polish)*. PhD thesis, Wroclaw University of Technology, Poland, 2008.

[12] M. F. Porter. An algorithm for suffix stripping. pages 313–316, 1997.

[13] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[14] Andrzej Sluzek. Building local features from pattern-based approximations of patches: Discussion onmoments and hough transform. *EURASIP Journal on Image and Video Processing*, 2009.

[15] Michal Stanek, Bartosz Broda, Halina Kwasnicka, and Mariusz Paradowski. Magma - efficient method for image annotation in low dimensional feature space based on multivariate gaussian models. In *In Proc. of IMCSIT 2009 (accepted)*, 2009.

[16] Ellen M. Voorhees and Lori P. Buckland, editors. *Proceedings of The Seventeenth Text REtrieval Conference, TREC 2008, Gaithersburg, Maryland, USA, November 18-21, 2008*, volume Special Publication 500-277. National Institute of Standards and Technology (NIST), 2008.

[17] James Z. Wang, Jia Li, and Gio Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001.