I2R AT IMAGECLEF WIKIPEDIA RETRIEVAL 2010

Kong-Wah WAN, Yan-Tao ZHENG, Sujoy ROY
Computer Vision and Image Understanding,
Institute for Infocomm Research,
1 Fusionopolis Way, Singapore 138632

Abstract

We report on our approaches and methods for the ImageCLEF 2010 Wikipedia image retrieval task. A distinctive feature of this year's image collection is that images are associated with unstructured and noisy textual annoations in three languages: English, French and German. Hence, besides following conventional text-based and multimodal approaches, we also focus some effort into investigating multilingual methods. We submitted a total of six runs along the following three directions: 1. augmenting basic text-based indexing with feature selection (three runs), 2. multimodal retrieval that re-ranks text-based results using visual-near-duplicates (VND), (one run), and 3. multilingual fusion that combines results from the three language resources indexed separately (two runs). Our best result (*i2rcviu MONOLINGUAL*, MAP of 0.2126) comes from the latter multilingual fusion approach, indicating the promise of exploiting multilingual resources. For our multimodal re-ranking run, we adopt a pseudo-relevance-feedback approach that builds a visual prototype model of each query without the need for any labeled example images. Essentially, we assume that the top-ranked image results from a text baseline retrieval are correct, and proceed to re-rank the result list such that images that are visually similar images to the top-ranked images are pushed up the ranks. This VND-based re-ranking is applied on the results of a text baseline (RUN *i2rcviu I2R.baseline*, MAP of 0.1847) that indexed images using all available annotations. This visual re-ranking run (*i2rcviu I2R.VISUAL.NDK*) achieves a MAP of 0.1984, a 7% improvement. Led by this encouraging result, we apply our VND re-ranking on the results from the multilingual run, and obtain our best retrieval result (not submitted) of 0.2338.

Keywords: Multimodal Retrieval, Visual re-ranking, Multilingual fusion

1 Introduction

We present our approach and methods in the Wikipedia-MM task of ImageCLEF 2010 [9]. In this year, a key distinctive feature of the benchmark image collection is that images are annotated with unstructured and noisy text in three languages: English (EN), French (FR), and German (DE). Hence, apart from conventional text-based and multimodal (visual+text) approaches, we investigated into ways to exploit the multilingual nature of the image corpus. We submitted a total of six runs, focusing our effort along the following three main directions.

1.1 Text-based Retrieval

Firstly, we explore feature selection and relevance feedback techniques to enhance text-based retrieval. Of our six submission runs, three are from this line of research. Our motivation is that text methods would continue to be the main contributor to accurate image retrieval. Hence, attempting to improve text-based methods would naturally form the bulk of our effort.

1.2 Multimodal Methods

Our second focus is on multimodal approach. Specifically, we take the return results of a text-based baseline, makes the assumption that most of these return results are relevant and correct, and proceed to analyse the top images for visual-near-duplicates (VND). VND images are then clustered and re-ranked so that they become closer in ranks. This has the effect of improving the ranks of lower-ranked images that are similar to higher-ranked images. We note instead of making the weak assumption that the top images are relevant, we could have used the example images that accompany the queries. Nonetheless we adopt our present method for the following reasons:

1. several reseachers [1] have already explored the use of the example query images to build a visual prototype model for the query topic;
2. in real world scenarios, users are unlikely to provide example query images, and
3. we aim to explore fully automatic methods.

Our approach is also closer to the spirit of pseudo relevance feedback (PRF) commonly used in the text community. Because of the lack of time, we have only submitted one run for this line of research. In this run, we apply the VND re-ranking method on the results from a simple text baseline. This text baseline indexes the images based on whatever textual annotations that are present in the XML metadata, ignoring whether those annotations are in EN, DE or FR.

1.3 Exploiting Multilingual Resources

Finally, our third focus, which we believe to be the most novel, is on exploiting multilingual cues. This year's benchmark image collection offers the unique opportunity for us to examine the comparative advantage of using multiple language resources on image retrieval.

Specifically, we build several image indexes based on the various combinations of annotation languages. For example, we build three image indexes based on a single annotation language, i.e. EN-only, DE-only and FR-only. For this case, if an image only has one annotation language, say, EN, then its DE and FR index will be empty, and a query issued to the DE index or FR index will not return that image. We also build three other image indexes based on two annotation languages, i.e. EN-DE, EN-FR, DE-FR. For this case, if an image has only EN annotation, then to build the EN-DE dual-lingual index, we will perform machine translation of EN to DE. For lack of time and compute resource, we stop short of building a triple-lingual index [1].

Because all queries are described using all three languages, for each query, we issue to an image index using the query text in the appropriate language. For example, we issue to the EN-only-index with the EN-only query text, to the DE-only-index the DE-only query text, and so on. Similarly for

[1]Note also that our default baseline text system is actually *NOT* a triple-lingual index system. It merely uses whatever annotations that is present. About 60% of images have annotations in only a single language, and 25% of images have annotations in two languages. Only 10% of images have annotations in all three language

the dual-lingual index, we form a concatenated new query text comprising description text in the respective languages. For example, we issue to the EN-DE-index the concatenated English query description and German query description.

Results from multiple image indexes can be fused by taking the maximum retrieval confidence. Clearly, there are configurable options to decide which image index to fuse from. We submitted a total of two runs: (1). one that fuses results from the three mono-lingual image index, and (2). one that fuses results from the three dual-lingual image index. The official evaluation results show that fusing results from the mono-lingual image index produces better results than that from fusing results from the dual-lingual image index.

The rest of the paper is organized as follows. Section 2 provides details of our methods and runs, with results from the official evaluation. Section 3 discusses the results from our submitted runs and presents some post-evaluation results from our further experimentations. Section 4 concludes the paper with some future outlook.

2 Our Methods and Results

2.1 Text-based retrievals

In all our experiments, we use the Lucene toolkit [2] as the retrieval system. In parsing the XML metadata files, we removed all mark-up tags, and retained the main textual data after stopword removal as the annotation content. To build the language-specific index, the annotation content is further splitted into its respective languages. The actual query text that is issued to a retrieval system is a concatenation of the query in its respective languages. Except for the run where we are exploring with query expansion, there is no other manipulation of the query terms.

2.1.1 Baseline Text Retrieval System – RUN: I2R.baseline

As mentioned earlier, the index for our baseline text retrieval system is built by utilizing whatever textual annotations that are present in the XML metadata files. It ignores the language of the annotations. A query is composed by concatenating the query description text in all three languages. This submitted run is called *i2rcviu I2R.baseline*. It obtained a MAP of 0.1847. Figure 1 shows the official result for this run.

	Participant	Run	Modality	FB/QE	Annotation language	Topic language	Topic field (s)	MAP	P@10	P@20	R-prec.	Bpref	NDCG	retrieved images	relevant retrieved images
43	i2rcviu	I2R.baseline	Textual	NOFB	EN+FR+DE	EN+FR+DE	TITLE	0.1847	0.4214	0.3936	0.2642	0.2278	0.4471	68247	6036

Figure 1: Results for our text baseline RUN:i2rcviu I2R.baseline

Compared to the best obtained results, this MAP value is not impressive. One likely reason for the suboptimal retrieval lies in the way we build this text baseline, and the way we issue concatenated queries of all three languages. Because only 10% of images have annotations in all three languages, for 90% of the images, there will be some terms in the concatenated queries that are non-informative. In other words, for the huge majority of images, there is a language mismatch. This results in noisy retrieval. However, we also note that this situation may be the norm in practical real-world scenarios, where there may not be enough resource to create separate indexes for each of the annotation language. Hence this baseline result can serve as a reference for the multilingual retrieval community.

2.1.2 Feature Selection – RUN: I2R.Feat.selection

Feature selection is a process wherein a subset of the text features (words) are selected for the final text vector representation. The main idea is to discard unimportant, non-informative words, and to retain a smaller subset of words that contribute the most to accurate retrieval [3, 4].

Our feature selection proceeds as follow. First, we build a new corpus by issuing to our baseline triple-lingual index the fully concatenated query description. For each query, we collate the top 1000 results. This means that we have a new corpus that is of maximum size 70K. We then perform feature selection on this new corpus. We apply a ad-hoc combination of feature selection techniques from [3, 4]. Specifically, we use the *Term Contribution* and *Document Frequency* metric to weight each unique term in the new corpus. We then sort the weighted terms and remove the top 15% and bottom 20% terms. As an extra check, for these words that are earmarked for removal, we further compute their ESA semantic relatedness [5] with every word in the query text of all 70 queries. If the ESA value of a word is above a threshold, we shall retain it. The idea is to avoid removing words that turn out to be individually relevant to a particular query. Using the new terms in the final list after removal, we create a new index. This submitted run is called *i2rcviu I2R.Feat.selection*. It obtained a MAP of 0.1945. Figure 2 shows the official result for this run.

	Participant	Run	Modality	FB/QE	Annotation language	Topic language	Topic field (s)	MAP	P@10	P@20	R-prec.	Bpref	NDCG	retrieved images	relevant retrieved images
32	i2rcviu	I2R.Feat.selection	Textual	NOFB	EN+FR+DE	EN+FR+DE	TITLE	0.1945	0.4600	0.4186	0.2684	0.2365	0.4561	68252	6216

Figure 2: Results for our text feature selection RUN:i2rcviu I2R.Feat.selection

Compared to the I2R.Feat.selection baseline run, there is a marginal improvement of 5%. This shows the utility of feature selection methods in text retrieval.

2.1.3 Query Expansion – RUN: I2R.PRF

We next experimented with a query expansion approach. We adopt Rocchio's pseudo-relevance feedback [6] as our query expansion model. By assuming the top return results to be relevant, the query expansion model reformulates the query by augmenting the original query with new feedback words selected from the top return results. For the sake of comparison, we also use the same *Term Contribution*, *Document Frequency* and *ESA* metrics to weight words in the top return documents. Amongst the top Term-Contribution and Document-Frequency words, we take the top K words with high ESA values, where K is the length of the original issued query. Hence the expanded query is of length $2K$. The submitted run is called *I2R.PRF*. It obtained a MAP of 0.1840. Figure 3 shows the official result for this run.

	Participant	Run	Modality	FB/QE	Annotation language	Topic language	Topic field (s)	MAP	P@10	P@20	R-prec.	Bpref	NDCG	retrieved images	relevant retrieved images
44	i2rcviu	I2R.PRF	Textual	FB	EN+FR+DE	EN+FR+DE	TITLE	0.1840	0.4386	0.3993	0.2656	0.2299	0.4388	68920	5895

Figure 3: Results for our pseudo relevance feedback-based qeury expansion RUN:i2rcviu I2R.PRF

This is a disappointing result. Not only is this worse than the Feature Selection run (I2R.Feat.selection), it is even worse than the baseline run (I2R.baseline). While the utility of query expansion has been

proven in many information retrieval tasks, we did not see its success generalize onto the present Wikipedia image retrieval task.

2.2 Multimodal Approach – RUN : I2R.VISUAL.NDK

Our multimodal strategy to image retrieval continues on the trend of combining visual processing with the results from text analysis. Specifically, we adopt a re-ranking approach that reorders images according to their visual similarity with a set of visual prototype models constructed from the top-ranked images. The main intuition of our method is that for a given query, and a visual model of that query, images that are visually closer to the visual model of the query, are likely to be more relevant to the query.

Given a visual model of a query, we use a visual-near-duplicate (VND) approach to compute visual similarity between an image and the visual model. Near-duplicate images denote a group of images that depict the same or duplicate scene in the whole or part of the image but with slightly varying visual appearance. The reason for visual difference is due to geometric, photometric and scale changes caused by the variance of camera shooting angle, lighting condition, camera sensor or photo editing process. If a group of near-duplicate images are returned for certain queries, this indicates a good probability that all of them are positive answers. Figure 4 below shows a few examples of near-duplicate examples that are both positive answers to the given queries.

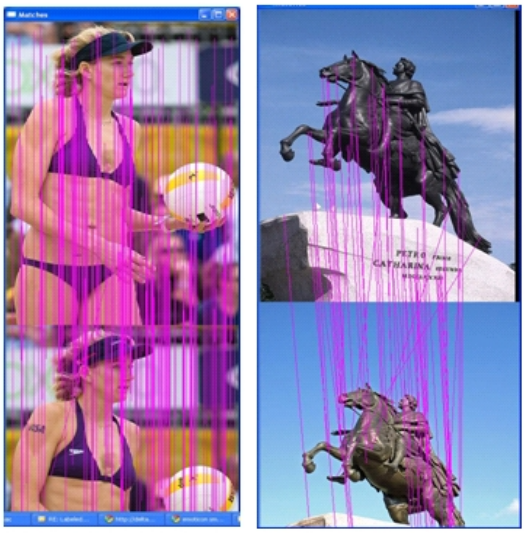

Figure 4: Example VND matching

There are many possible methods to construct a visual model of the query. A common strategy is to learn the model from a set of relevant images returned from an image search engine. For the set of official queries in WikipediaMM 2010, each query comes with three example images to visually illustrate the query intent. We note that some researchers have already applied learning a visual model of the query from these images [1]. However, the downside of this approach is that because the results of an image search could be noisy, there needs to be some manual effort involved in ascertaining the relevance of the returned images. This is especially problematic for non-object type of queries, where returned images tend to be even noiser.

In this paper, we experimented with a pseudo relevance feedback approach to build a visual model of the query by using the top-ranked images. The implicit assumption is that these top-

ranked images are likely to be relevant. We choose the text baseline run (*I2R.baseline*) on which we will apply the visual re-ranking model. We follow the image-cluster-matching approach in [7] to build a visual model of each query from a set of V_{pos} images, comprising the top-20-ranked images by the text baseline run *I2R.baseline*.

We use a local-feature-based representation for images. Each image is first computed for a number of key-points and their descriptors. The similarity between two images is then determined by matching their keypoint descriptors. Specifically, we utilize Difference of Gaussian as keypoint detector and Scale Invariant Feature Transform (SIFT) as local descriptor. After identifying the group of near-duplicate images in the retrieval list, we take a simple heuristic to rank them at the top as follow. For each image i, we sum up its distance D_i to the each image in V_{pos} as $D_i = (\sum_{j \in V_{pos}, j \neq i} distance(i,j))/|V_{pos} - \{i\}|$. We implement the $distance(.)$ function as a variant of the keypoint-based matching method in [8]. Our submitted run for this method is called *I2R.VISUAL.NDK*. It obtained a MAP of 0.1984. Figure 5 shows the official result for this run.

	Participant	Run	Modality	FB/QE	Annotation language	Topic language	Topic field (s)	MAP	P@10	P@20	R-prec.	Bpref	NDCG	retrieved images	relevant retrieved images
29	i2rcviu	I2R.VISUAL.NDK	Mixed	NOFB	EN+FR+DE	EN+FR+DE	TITLE	0.1984	0.4971	0.4321	0.2761	0.2466	0.4546	64830	6084

Figure 5: Results for our visual re-ranking method RUN:i2rcviu I2R.VISUAL.NDK

From the MAP values of 0.1984, we see that there is an improvement of 7% on the text baseline (MAP 0.1847). This is a healthy sign, and points to the effectiveness of using our visual re-ranking model. Note that this improvement comes about without the need for any training images or manual labeling effort. We report in section 3 further experiments that confirms the ability of our visual re-ranking model to improve results from other text retrieval baselines.

2.3 Exploiting Multilingual Cues

In this year, we are provided with an image collection that is annotated with text in three languages: English, German and French. This offers us a good opportunity to explore the utility of these multilingual resources for improved image retrieval. Intuitively, the additional text annotations should help in retrieval since the information are highly related, but not in a redundant way. For example, if an "Volcano" image has the English word "Volcano", then obviously it would match well to the English query with the word "Volcano". If the same image does not have an English annotation, but rather it has a German equivalent translation, then it would still likely match well to the original English query translated to the German "Vulkan". Further more, if the image has both English and German annotations, then a query issued with both "Volcano" and "Vulkan" would have a more confident hit on this image.

Of course in a real-world setting, not all images would come with multilingual annotations. In fact, the statistics of the 2010 WikipediaMM collection is such that 60% of the images have annotations in only a single language, 25% of images have annotations in two languages, and only 10% of images have annotations in all three language. This means that the huge majority of images are single-language-annotated. In terms of medium of language in the annotations, 60% of the images have annotations in English, followed by 46% of images having annotations in German, and 30% of images having annotations in French. In such a situation, we explore the following questions:

1. Will we do better if we create a separate mono-lingual index, and issue to it with the appropriate query text in the corresponding language?
2. Will we do better if we create a separate dual-lingual index, and issue to it with the appropriately concatenated query text in the corresponding languages?

Each of the above is compared against the default text baseline used in our paper: the *I2R.baseline* system, which builds a generic index that uses all the text annotations that are present in the XML metadata, disregarding which language the text are in. We report our exploration of both question and their results in the following.

2.3.1 Mono-lingual – RUN: MONOLINGUAL

The main idea is simple. Since the bulk of images would have annotations in at least one language, and since all our queries have the equivalent translations in all three languages, we can build a new retrieval system by the following:

1. build a mono-lingual index for all images
2. issue the appropriate translated-queries to the corresponding mono-lingual index
3. fuse the three result lists using maximum confidence

Note that in building the language-specific index, we *do not* require that all images must have all three annotation languages. In contrast to the dual-lingual run in the next subsection, we do not perform any machine translation here. In other words, if an image has only EN annotation, but no annotation in DE nor in FR, then only the EN-index contains this image, and the DE-index and the FR-index will not contain this image. During query time, only the English query description text will be issued to the EN-index.

	Participant	Run	Modality	FB/QE	Annotation language	Topic language	Topic field (s)	MAP	P@10	P@20	R-prec.	Bpref	NDCG	retrieved images	relevant retrieved images
16	i2rcviu	MONOLINGUAL	Textual	NOFB	EN	EN+FR+DE	TITLE	0.2126	0.4486	0.4143	0.2832	0.2585	0.4805	68739	6955

Figure 6: Results for our mono-lingual RUN:i2rcviu MONOLINGUAL

Our submitted run for this method is called *MONOLINGUAL*. It obtained a MAP of 0.2126. Out of our total six submitted runs, this run has the best result. Compared to the MAP of 0.1847 from the baseline run (*I2R.baseline*), this is a big improvement of 15%. Note that because the baseline *I2R.baseline* run builds a generic index that combines all available annotations in all languages, the better result from *MONOLINGUAL* tells us that it is better to be language-specific. Figure 6 shows the official result for this run.

2.3.2 Dual-lingual – RUN: DUAL_LINGUAL

Encouraged by the promising result from our *MONOLINGUAL* run, we turn to the question of whether we can do better with dual-lingual index. We build a new retrieval system following similar ideas in *MONOLINGUAL*:

1. build a dual-lingual index for all images. Perform machine translation if neccessary.

2. issue the concatenation of the appropriate translated-queries to the corresponding dual-lingual index
3. fuse the three result lists using maximum confidence

Note that in building the dual-lingual index, we now mandate that all images must have all three annotation languages. In other words, if an image has only EN annotation, but no annotation in DE nor in FR, then we perform machine translation of the En-text to both DE-text and FR-text. We used the Google AJAX Translation API for the machine translation. Limited by the constraints in the AJAX web services, we were forced to throttle our calls, and only performed translation of the headline snippets of the Wikipedia text. Our submitted run for this method is called *DUAL_LINGUAL*. It obtained a MAP of 0.1742. Figure 7 shows the official result for this run.

	Participant	Run	Modality	FB/QE	Annotation language	Topic language	Topic field (s)	MAP	P@10	P@20	R-prec.	Bpref	NDCG	retrieved images	relevant retrieved images
56	i2rcviu	DUAL_LINGUAL	Textual	FB	EN+FR+DE	EN+FR+DE	TITLE	0.1742	0.4271	0.3964	0.2628	0.2263	0.4223	68072	5610

Figure 7: Results for our mono-lingual RUN:i2rcviu DUAL_LINGUAL

The result is disappointing. It is nowhere near to the MAP value of 0.2126 of the *MONOLINGUAL* run. Compared to the MAP of 0.1847 from the baseline run (*I2R.baseline*), there is even a drop of 6%. We attribute the bad results to the heavily subdued translation effort by our throttled AJAX call to Google Translation service. It is also likely that the automatically detected headline snippets submitted for translation is not meaningful and representative of the main content.

3 Further Experimentations

Of the six submitted runs, there are two promising results, namely our visual re-ranking *I2R.VISUAL.NDK* run and the mono-lingual *MONOLINGUAL* run. The visual re-ranking has improved retrieval over the baseline result by 7%, while using the mono-lingual index approach, retrieval improved by 15% over the baseline. As part of our post-evaluation effort, we apply the visual re-ranking method onto the best run from the official evalaution, namely the *MONOLINGUAL* run. We obtained a MAP result of 0.2338, an improvement of about 10%. This shows that the improvement from our visual re-ranking is robust and generalizable. Had we decided to submit this additional run, this result would have been ranked 13^{th} amongst all submissions.

4 Conclusion

For our participation in the ImageCLEF 2010 Wikipedia retrieval task, we submitted a total of six runs, exploring the following directions: (1). Feature selection and feedback strategies text-based methods, (2). Visual re-ranking without the need for any labeled images, (3). Fusion of mutilingual resources. From the official evaluation results, we see that both our visual re-ranking method and the fusion of mono-lingual resources can significantly improve retrieval. Both strategies will now form an integral part of our future effort in image retrieval.

References

[1] Debora Myoupo, Adrian Popescuy, Herve Le Borgne and Pierre-Alain Moellic, "Visual Reranking for Image Retrieval over the Wikipedia Corpus", ImageCLEF 2009 working notes, 2009.

[2] Lucene, http://lucene.apache.org/java/docs/.

[3] Tao Liu, Shengping Liu and Zheng Chen, "An Evaluation on Feature Selection for Text Clustering", In Proc ICML, pp 488-495, 2003.

[4] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization", In Proc of ICML, pp 412-420, 1997.

[5] Evgeniy Gabrilovich and Shaul Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis", In Proc International Joint Conference on Artificial Intelligence, pp 1606-1611, 2007.

[6] J. Rocchio, "Relevance feedback in information retrieval", In Gerard Salton, editor, The SMART Retrieval System – Experiments in Automatic Document Processing, pp 313-323, 1971.

[7] O. Boiman, E. Shechtman, and M. Irani, "In Defense of Nearest-Neighbor Based Image Classification", In Proc CVPR, 2008.

[8] Y. Ke, R. Sukthankar, and L. Huston, "An efficient parts-based near-duplicate and sub-image retrieval system", In Proc ACM Multimedia, pp 869-876, 2004.

[9] Adrian Popescu, Theodora Tsikrika and Jana Kludas, "Overview of the Wikipedia Retrieval task at ImageCLEF 2010", In the Working Notes of CLEF 2010, 2010.