DEMIR at ImageCLEFMed 2011: Evaluation of Fusion Techniques for Multimodal Content-based Medical Image Retrieval

Adil Alpkocak, Okan Ozturkmenoglu, Tolga Berber,
Ali Hosseinzadeh Vahid and Roghaiyeh Gachpaz Hamed

Dokuz Eylul University
Department of Computer Engineering
DEMIR Dokuz Eylul Multimedia Information Retrieval Research Group
Tinaztepe, 35160 Izmir, Turkey
alpkocak@cs.deu.edu.tr, okan.ozturkmenoglu@deu.edu.tr, tberber@cs.deu.edu.tr,
{ali_h_vahid, ramisa_84}@yahoo.com

Abstract. This paper present the details of participation of DEMIR (Dokuz Eylul University Multimedia Information Retrieval) research team to the context of our participation to the ImageCLEF 2011 Medical Retrieval task. This year, we evaluated fusion and re-ranking method which is based on the best low level feature of images with best text retrieval result. We improved results by examination of different weighting models for retrieved text data and low level features. We tested multi–modality image retrieval in ImageCLEF 2011 medical retrieval task and obtained the best seven ranks in mixed retrieval, which includes textual and visual modalities. The results clearly show that proper fusion of different modalities improve the overall retrieval performance.

Keywords: Information Retrieval, Weighting-schemes, Re-ranking, Medical Imaging, Content-based Image Retrieval, Medical Image Retrieval.

1 Introduction

In this paper we present the experiments performed by Dokuz Eylul University Multimedia Information Retrieval (DEMIR) Group, Turkey, in the context of our participation to the ImageCLEF 2011 Medical Image retrieval task [1]. The main focus of this work is to improve results by evaluation of different weighting models in text retrieval and then choose the best low-level feature of images for fusion with text only results. During the combination of text and low-level features, we check the variation of methods to gain the best result. On the other hand, we performed the experiments for narrowing down the data collection by defining and filtering out of irrelevant documents. Also we checked the weighted querying system performance in retrieval systems by weighting the special words in queries.

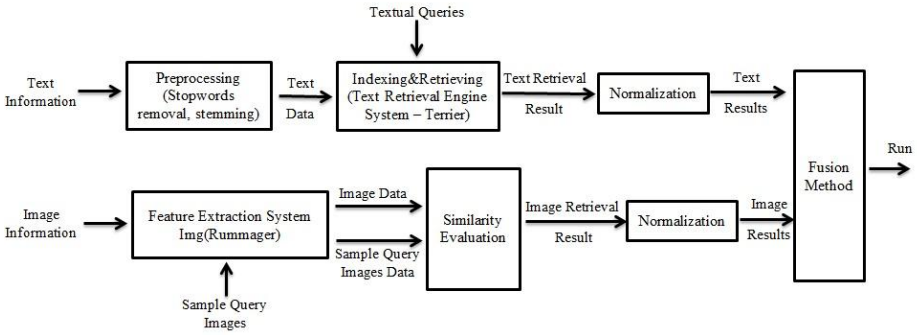

Fig. 1. Basic block diagram of retrieval system.

After analyze the visual and textual features set we used (Section 2), we describe the multimodal fusion techniques for multimodal information (Section 3). After we present experiments on ImageCLEF 2010 Medical and Wikipedia Retrieval tracks data (Section 4), then Section 5 concludes the paper by pointing out the open issues and possible avenues of further research in the area of multimodal re-ranking and fusion techniques for content-based image retrieval.

2 The Feature Set

The data collection of ImageCLEF 2011 Medical retrieval has textual and visual information. Participants will be given a set of 30 textual queries with 2-3 sample images for each query. The queries will be classified into textual, visual and mixed, based on the methods that are expected to yield the best results.[1]

We performed our experiments using ImageCLEF 2010 Medical and Wikipedia Retrieval track's text and image data. We check the variation of retrieval methods on textual and visual information to gain the best result.

2.1 Textual Features

Since the choice of the weighting model may crucially affect the performance of any information retrieval system, first of all we decided to work on evaluating the relative merits and drawbacks of different weighting models using Terrier IR Platform [2], open source search engine written in Java and is developed at the School of Computing Science, University of Glasgow.

We performed our experiments on textual features using ImageCLEF 2010 Medical track collection. We started from a traditional bag-of-words representation of pre-processed texts that pre-processing includes stemming (Porter stemmer [3] for English) and stop words removal. DFR- BM25 model's MAP score is not the best one, but the all weighting model's number of relevant retrieved score results are close to each other and considering achievements of this model [11], we submitted our

textual base point run using this model on ImageCLEF 2011 Medical retrieval task data collection as RUN_1 .

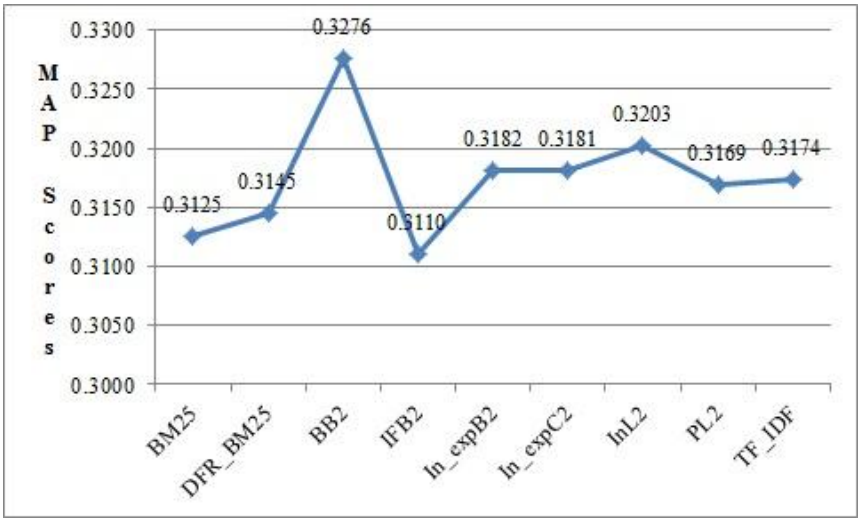

Fig. 2. MAP scores of weighting models for textual features

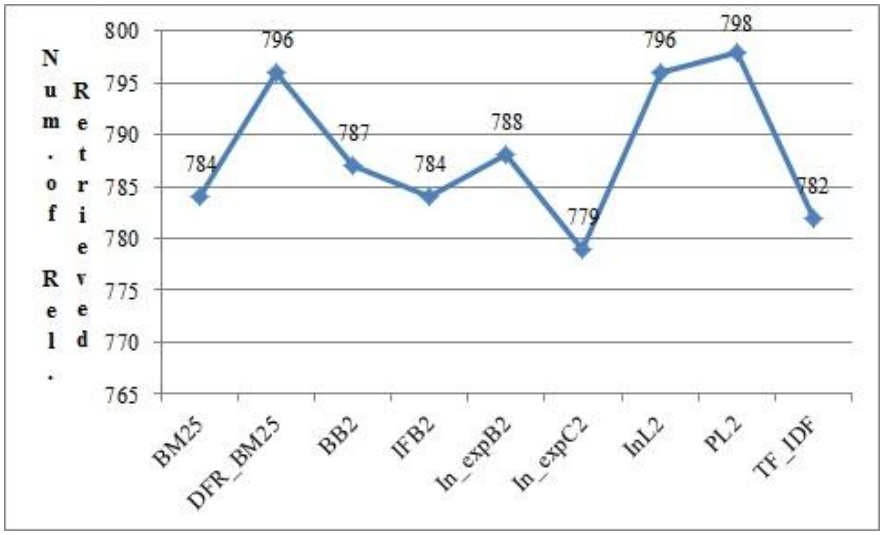

Fig. 3. Number of relevant retrieved document in different weighting models for textual features

2.2 Visual Features We Used

Selection of low-level features is one of the major aspects of a typical content-based information retrieval (CBIR) system. We call these low-level features because most

of them are extracted directly from digital representations of objects in the database and have little or nothing to do with human perception. Thanks to Img(Rummager) application [4], is developed in the Automatic Control Systems & Robotics Laboratory at the Democritus University of Thrace-Greece, and we extracted features explained below for all images in ImageCLEF 2011 test collection and query examples:

- **EHD:** This Edge Histogram Descriptor proposed for MPEG-7 expresses only the local edge distribution in the image and is designed to contain only 80 bins for this purpose. The EHD basically represents the distribution of 5 types of edges in each local area called a sub-image that is defined by dividing the image space into 4x4 non-overlapping blocks. Thus, the image partition always yields 16 equal-sized sub-images regardless of the size of the original image. Edges in the sub-images are categorized into 5 types: vertical, horizontal, 45-degree diagonal, 135-degree diagonal and non-directional edges. Thus, the histogram for each sub-image represents the relative frequency of occurrence of the 5 types of edges in the corresponding sub-image and contains 5 bins [7].
- **CEDD:** This feature combines EHD with color histogram information and named "Color and Edge Directivity Descriptor". CEDD size is limited to 54 bytes per image, rendering this descriptor suitable for use in large image databases. Important attribute of the CEDD is the low computational power needed for its extraction, in comparison to the needs of the most MPEG-7 descriptors [4].
- **FCTH:** This feature fuzzy version of CEDD feature which contains fuzzy set of color and texture histogram and named "Fuzzy Color and Texture Histogram". This feature contains results from the combination of 3 fuzzy systems including histogram, color and texture information. FCTH size is limited to 72 bytes per image, and also suitable for use in large image databases [5].
- **BTDH:** This feature is very similar to FCTH feature. The main difference from FCTH feature is using brightness instead of color histogram. This feature is originally developed for radiology images which do not contain color data [6].

After extracting features, we gain an *n*-dimensional feature space per feature. For query processing, we had to map all of the objects in the database and the query onto this space and then evaluate the similarity difference between the vector corresponding to the query and the vectors representing the data. We selected the Euclidean distance, one of commonly used similarity and distance functions for measuring distances between points in the 3D space, as distance/similarity function and based on obtained similarity scores; we found that CEDD and FCTH are the best descriptors for image retrieval based on low level features only. Therefore we submitted our visual only base point run for CEDD feature. Moreover we use these features for multimodal fusion in next experiments.

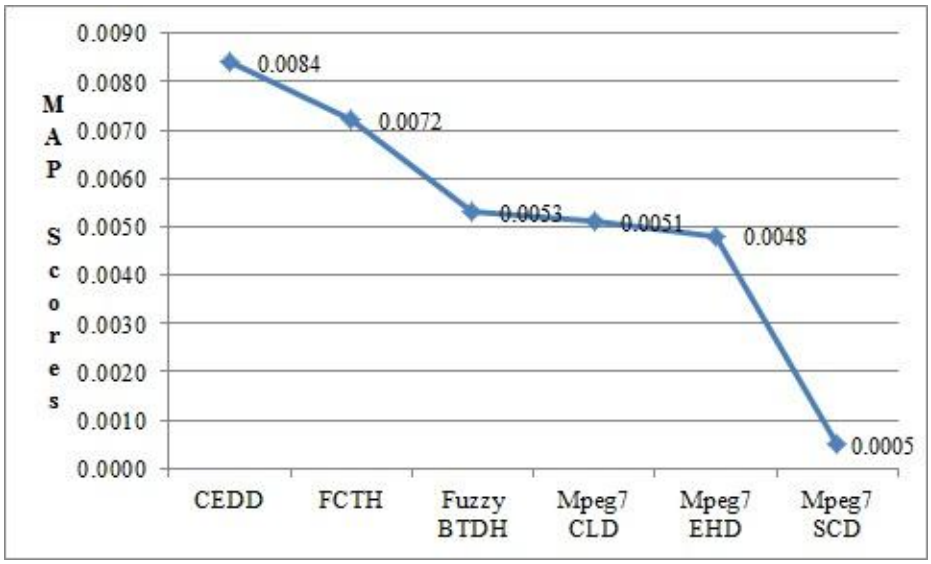

Fig. 4. Comparison of low level feature performance on ImageCLEF 2010 Wikipedia Retrieval task.

3 Fusion Techniques in Multimodal Information Retrieval

Multimedia fusion is referred to as integration of multiple media, their associated features, or the intermediate decisions in order to perform an analysis task, has gained much attention of many researchers in recent times. The fusion of multiple modalities can provide complementary information and increase the accuracy of the overall decision making process [8].

The fusion of different modalities is generally performed at two levels: feature level or early fusion and decision level or late fusion. Some researchers have also followed a hybrid approach by performing fusion at the feature as well as the decision level. In the feature level or early fusion approach, the features, some distinguishable properties of a media stream, extracted from input data are first combined and then sent as input to a single analysis unit that performs the analysis task. In the decision level or late fusion approach, the analysis units first provide the local decisions D_1 to D_n that are obtained based on individual features F_1 to F_n. Then a decision fusion unit combines local decisions to make a fused decision vector that is analyzed further to obtain a final decision D about the task or the hypothesis. To achievement the advantages of both the feature level and the decision level fusion strategies, several researchers have opted to use a hybrid fusion strategy, which is a combination of both feature and decision level strategies.

The decision level fusion strategy has many advantages over feature fusion. For instance, the decisions (at the semantic level) usually have the same representation. Therefore, the fusion of decisions becomes easier. Moreover, the decision level fusion strategy offers scalability (i.e. graceful upgrading or degradation) in terms of the

modalities used in the fusion process, which is difficult to achieve in the feature level fusion. Another advantage of late fusion strategy is that it allows us to use the most suitable methods for analyzing each single modality and this provides much more flexibility than the early fusion.

Because of these profits, we exerted Linear Weighted Fusion, one of the simplest and most widely used methods on our extracted CEDD and FCTH similarity scores and similarity scores that gained from text retrieval as explained in previous chapters. We applied Fagin's Combination Algorithms [9] for Ranked Input Sets putting on two score aggregation function defined as "Average" and "Weighted Average". The average function is applied by taking mean of individual similarity scores of any object.

On the other hand, the weighted average function is applied in the same manner but differing on multiplying each individual similarity with a weight value. The weight assignment to individual scores provides an importance level for each feature defined in a whole query [10]. After comparison of several studies we decided to multiply textual feature by 3 and CEDD feature by 2 to gain the best fusion result based on weighted average combination method.

Before fusion operation takes place, normalization should be applied to get accurate and correct results since different modalities results a different ranges of similarity values [12]. Here, we applied Min-Max normalization on similarity values to ensure that the range of these features is between 0 and 1. The following equations will ensure the range of this feature from 0 to 1.

Suppose the range for a feature x_i is from x_{min} to x_{max}. Then the normalized feature x_i' is defined as follows:

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Min-Max normalization is a process of taking data measured in its units and transforming it to a value between 0.0 and 1.0. The lowest (min) value is set to 0.0 and the highest (max) value is set to 1.0. This provides an easy way to compare values that are measured using different scales (i.e., textual, shape, visual, density etc.) or different units of measure (i.e., Euclidean or non-metric space values). After normalization of the similarity values, we combined the different modalities in ranked results.

4 Experimentations

We submitted 10 runs to ImageCLEF Medical Retrieval task, in three different categories. The first category includes the runs for baseline retrieval in single modality, numbered as 1 and 3 are baseline retrievals for textual-only and visual-only retrieval, respectively. The second groups of runs to evaluate re-ranking affects to base line, numbered as 8 is re-indexed the baseline retrieval result and re-ranked in textual modality. The last group includes mixed retrieval experiments with fusion of different modalities, numbered as 2, 4, 5, 6, 7, 9 and 10. As illustrated in Table 1, it is obvious that results of mixed runs are better than textual or visual only. Moreover

results of weighted average combination method are better than normal average method in all approach. Below, we provide a short description of each runs, shortly.

- **RUN_1**: This run is our baseline retrieval result for textual modality. In this run, we removed the stop words, applied Porter stemmer algorithm and used the DFR_BM25 weighting model on text retrieval engine system, Terrier. Let the subscript indicates the arbitrary run ID, the similarity of first run, S_1, is defined as follows:

$$S_1 = text\_baseline\_score \quad (2)$$

- **RUN_3**: Our baseline retrieval result is this run for visual modality. We used the CEDD feature in visual modality because its performance is better than other features, also you can see in Figure 4.

$$S_3 = visual\_baseline\_cedd\_score \quad (3)$$

- **RUN_8**: This run of our group on textual feature is based on our proposed a two-level re-ranking approach in for move relevant documents upward. Re-ranking is a method to reorder the initially retrieved documents with the aim to increase precision. Basically, relevant documents with low similarity scores are re-weighted and reordered. In this run, we propose a new re-ranking approach which includes the narrowing-down phase of search space. Result sets of each query and corresponding base similarity scores are inputs for re-ranking operation. Firstly, we selected relevant documents using initial similarity scores. In other word, we filtered out non-relevant documents based on initial similarity scores. For this we selected first 1000 relevant documents if it existed. Then we constructed a new VSM using this small document sets. This operation drastically reduced both the number of documents and the number of terms. In short, this level shrinks down the initial VSM data into more manageable size. Then we calculated similarity score of new VSM and submitted the results as RUN_8. As illustrated in Table 1, unlike the achievements of this approach in ImageCLEF 2010 Wikipedia retrieval task, all factors of retrieval system decline in contrast to our textual base line run.

$$S_8 = rerank\_reindex\_text\_score \quad (4)$$

- **RUN_2:** Another narrowing down approach that we examine this year is based on medical image modality classification. Result sets of each query and corresponding base similarity scores and their class based on any classification algorithm are inputs for this approach. We also expanded query structure by assignment a type for example images of each query. A query can have a more than one type. In the narrowing down phase we filtered out non relevant images that its class was not the same as corresponding query type. We applied this method filtering the modality classification using GIFT system and 1NN approach and submitted RUN_2 as results. As obtained from Table 1, although MAP in this method is decreased but there are a considerable improvement in P@10 and P@ 20 values in contrast to textual base line.

$$S_2 = (text\_baseline\_score + filtered\_out\_by\_modality\_classification)/2 \quad (5)$$

- **RUN_5**: In this run, we combined the multiplied textual feature by 3 with the multiplied visual retrieval result using CEDD feature by 2, divided total score with the rated value 5.

$$S_5 = (3 \times text\_baseline\_score + 2 \times cedd\_score)/5 \quad (6)$$

- **RUN_4**: We combined the baseline textual retrieval result with visual retrieval result using CEDD feature and get average score.

$$S_4 = (text\_baseline\_score + cedd\_score)/2 \quad (7)$$

- **RUN_7**: Another approach that we experimented in text retrieval this year is evaluation of effects of weighting to special words in queries. For this purpose we selected the medical modality names in queries (i.e., CT, PET, X-RAY, MRI etc.) and weighted them by 2.5 using query language of Terrier. Although result of this approach decline in compare to baseline too, but they are better than result of re-ranking methods. Due to limitation of submitted runs of participant, we did not submitted weighted text retrieval results as a new run but we fused them with low level feature of images to obtain better performance. In this run, we combined the multiplied weighted textual feature by 3 with the multiplied visual retrieval result using CEDD feature by 2, divided total score with the rated value 5.

$$S_7 = (3 \times weighted\_text\_score + 2 \times cedd\_score)/5 \quad (8)$$

- **RUN_10**: After we combined the multiplied RUN_8 result by 3 with the multiplied visual retrieval result using CEDD feature by 2 and divided total score with the rated value 5.

$$S_{10} = (3 \times \text{rerank\_reindex}\_text\_score + 2 \times cedd\_score)/5 \quad (9)$$

- **RUN_6:** After we combined the weighted textual retrieval result with visual retrieval result using CEDD feature and get average score.

$$S_6 = (weighted\_text\_score + cedd\_score)/2 \quad (10)$$

RUN_9: After we combined the RUN_8 result with visual retrieval result using CEDD feature and get average score.

$$S_9 = (rerank\_reindex\_text\_score + cedd\_score)/2 \quad (11)$$

Table 1. Runs of DEMIR group in ImageCLEFMed 2011.

RunID	Rank	Type	MAP	P10	P20	Rprec	bpref	rel_ret
5	1	Mixed	0.2372	0.3933	0.3550	0.2881	0.2738	1597
4	2	Mixed	0.2307	0.3967	0.3400	0.2706	0.2606	1595
7	3	Mixed	0.2014	0.3400	0.3233	0.2587	0.2481	1455
10	4	Mixed	0.1983	0.4067	0.3350	0.2397	0.2428	1349
6	5	Mixed	0.1972	0.3367	0.3083	0.2489	0.2383	1443
9	6	Mixed	0.1853	0.3667	0.3283	0.2309	0.2230	1338
2	7	Mixed	0.1645	0.3967	0.3350	0.2340	0.2198	890
1	15	Text	0.1942	0.3400	0.2933	0.2242	0.2215	1444
8	49	Text	0.1452	0.3033	0.2633	0.1683	0.1859	1288
3	12	Visual	0.0174	0.1067	0.0833	0.0434	0.0602	569

5 Conclusion

In this year, we examined effects of different weighting models on text retrieval and found that the role of proper weighting model selection is to improve the performance of text retrieval systems. Also, we compare MAP of different extracted low-level features normalized similarity scores and due to this comparison we select CEDD and FCTH descriptors as suitable features to utilize for fusion to textual results. Also due to analogy of combination methods in our previous studies, we acquire choosing a suitable combination method for fusion improved the results. The results clearly show that combining text-based and content-based image retrieval results with a proper fusion technique improves the performance.

References

1. Medical Image Retrieval Task 2011, http://www.imageclef.org/2011/medical
2. The Terrier IR Platform, http://terrier.org/docs/v2.2.1/
3. Porter, M.F.: An algorithm for suffix stripping, Program: electronic library and information systems, vol. 14, iss. 3, pp. 130--137 (1980)
4. Chatzichristofis, S.A., Boutalis, Y.S., Lux, M.: Img(Rummager): An Interactive Content Based Image Retrieval System. In: 2nd International Workshop on

Similarity Search and Applications, pp. 151--153. IEEE Computer Society, Washington (2009)
5. Chatzichristofis, S.A., Boutalis, Y.S.: FCTH: Fuzzy Color and Texture Histogram - A Low Level Feature for Accurate Image Retrieval. In: 9th International Workshop on Image Analysis for Multimedia Interactive Services, vol., no., pp.191--196. Klagenfurt, Austria (2008)
6. Chatzichristofis, S.A., Boutalis Y.S.: Content based radiology image retrieval using a fuzzy rule based scalable composite descriptor. Multimedia Tools and Applications, vol. 46, iss. 2, pp. 493--519 (2009)
7. Won C. S., Park D. K., Park S.J.: Efficient Use of MPEG-7 Edge Histogram Descriptor, ETRI Journal, vol. 24, no. 1 (2002)
8. Pradeep K. Atrey, Anwar Hossain M.: Multimodal fusion for multimedia analysis, Multimedia Systems, vol 16, pp. 345--379 (2010)
9. Fagin R, Lotem A, Naor M.: Optimal aggregation algorithms for middleware, In: Journal of Computer and System Sciences, vol. 66, pp. 614--656 (2003)
10. Croft, W.B.: Combining Approaches to Information Retrieval, In: Advances in Information Retrieval, vol. 7, pp. 1--36 (2002)
11. He, Ben., Ounis, Iadh: Term Frequency Normalisation Tuning for BM25 and DFR Models., Advances in Information Retrieval, vol. 3408, pp. 200--214 (2005)
12. Ulker T.: Analysis and comparison of combination algorithms for joining ranked inputs, MSc Thesis, Dokuz Eylül University Department of Computer Engineering, Izmir, Turkey (2003)