Multi-Modal Interactive Approach to ImageCLEF 2007 Photographic and Medical Retrieval Tasks by CINDI

M. M. Rahman, Bipin C. Desai, Prabir Bhattacharya
Dept. of Computer Science & Software Engineering, Concordia University
1455 de Maisonneuve Blvd., Montreal, QC, H3G 1M8, Canada
mah_rahm@cs.concordia.ca

Abstract

This paper presents the contribution of CINDI group to the ImageCLEF 2007 ad-hoc retrieval tasks. We experiment with multi-modal (e.g., image and text) interaction and fusion approaches based on relevance feedback information for image retrieval tasks of photographic and medical image collections. For a text-based image search, keywords from the annotated files are extracted and indexed by employing the vector space model of information retrieval. For a content-based image search, various global, semi-global, region-specific and visual concept-based features are extracted at different levels of image abstraction. Based on relevance feedback information, multiple textual and visual query refinements are performed and user's perceived semantics are propagated from one modality to another with query expansion. The feedback information also dynamically adjusts intra and inter-modality weights in linear combination of similarity matching functions. Finally, the top ranked images are obtained by performing both sequential and simultaneous retrieval approaches. The analysis of results of different runs are reported in this paper.

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.3.7 Digital Libraries; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Object Recognition*

General Terms

Algorithms, Performance, Experimentation

Keywords

Content-based image retrieval, Vector space model, Feature extraction, Query expansion, Relevance feedback, Data fusion.

1 Introduction

For the 2007 ImageCLEF competition, CINDI research group has participated in two different tasks of ImageCLEF track: an ad-hoc retrieval from a photographic collection (e.g., IAPR data set) and ad-hoc retrieval from a medical collection (e.g., CASImage, MIR, PathoPic, Peir, endoscopic and myPACS data sets) [1, 2]. The goal of the ad-hoc task is given a multilingual statement describing a user information need along with example images, find as many relevant images as possible from the given collection. Our work exploits advantages of both text and image modalities by involving users in the retrieval loop for cross-modal interaction and integration. This paper presents our multi-modal retrieval methodologies, description of submitted runs, and analysis of retrieval results.

2 Text-Based Image Retrieval Approach

This section describes the text-based image retrieval approach where a user submits a query topic using keywords to retrieve images which are associated with retrieved annotation files. For a text-based search, it is necessary to prepare the document collection consisting of annotated XML and SGML files into an easily accessible representation. Each annotation file in the collection is linked to image(s) either in a one-to-one or one-to-many relationships. To incorporate a keyword-based search on these annotation files, we rely on the vector space model of information retrieval [3]. In this model, a document is represented as a vector of words where each word is a dimension in an Euclidean space. The indexing is performed by extracting keywords from selected elements of the XML and SGML documents depending on the image collection. Let, $T = \{t_1, t_2, \cdots, t_N\}$ denotes the set of keywords (terms) in the collection. A document D_j is represented as a vector in a N-dimensional space as $\mathbf{f}_{D_j} = [w_{j1}, \cdots, w_{jk}, \cdots, w_{jN}]^T$. The element $w_{jk} = L_{jk} * G_k$ denotes the *tf-idf* weight [3] of term $t_k, k \in \{1, \cdots, N\}$, in a document D_j. Here, the local weight is denoted as $L_{jk} = log(f_{jk}) + 1$, where f_{jk} is the frequency of occurrence of keyword t_k in a document D_j. The global weight G_k is denoted as inverse document frequency as $G_k = log(M/M_k)$, where M_k is the number of documents in which t_k is found and M is the total number of documents in the collection. A query D_q is also represented as an N-dimensional vector $\mathbf{f}_{D_q} = [w_{q1}, \cdots, w_{qk}, \cdots, w_{qN}]^T$. To compare D_q and D_j, the cosine similarity measure is applied as follows

$$\text{Sim}_{\text{text}}(D_q, D_j) = \frac{\sum_{k=1}^{N} w_{qk} * w_{jk}}{\sqrt{\sum_{k=1}^{N}(w_{qk})^2} * \sqrt{\sum_{k=1}^{N}(w_{jk})^2}} \tag{1}$$

where w_{qk} and w_{jk} are the weights of the term t_k in D_q and D_j respectively.

2.1 Textual Query Refinement by Relevance Feedback

Query reformulation is a standard technique for reducing ambiguity due to word mismatch problem in information retrieval [4]. In the present work, we investigate interactive way to generate multiple query representations and their integration in a similarity matching function by applying various relevance feedback methods. The relevance feedback technique prompts the user for feedback on retrieval results and then use that information on subsequent retrievals with the goal of increasing retrieval performance [4, 5]. We generate multiple query vectors by applying various relevance feedback methods. For the first method, we use the well known Rocchio algorithm [6] as follows

$$\mathbf{f}^m_{D_q}(Rocchio) = \alpha\ \mathbf{f}^o_{D_q} + \beta \frac{1}{|R|} \sum_{\mathbf{f}_{D_j} \in R} \mathbf{f}_{D_j} - \gamma \frac{1}{|\hat{R}|} \sum_{\hat{\mathbf{f}}_{D_j} \in \hat{R}} \hat{\mathbf{f}}_{D_j} \tag{2}$$

where $\mathbf{f}^m_{D_q}$ and $\mathbf{f}^o_{D_q}$ are the modified and the original query vectors, R and $\hat{R}$ are the set of relevant and irrelevant document vectors and α, β, and γ are weights. This algorithm generally moves a new query point toward relevant documents and away from irrelevant documents in feature space [6]. For our second feedback method, we use the *Ide-dec-hi* formula as

$$\mathbf{f}^m_{D_q}(Ide) = \alpha\ \mathbf{f}^o_{D_q} + \beta \sum_{\mathbf{f}_{D_j} \in R} \mathbf{f}_{D_j} - \gamma \max_{\hat{R}}(\mathbf{f}_{D_j}) \tag{3}$$

where $\max_{\hat{R}}(\mathbf{f}_{D_j})$ is a vector of the highest ranked non-relevant document. This is a modified version of the Rocchio's formula which eliminates the normalization for the number of relevant and non-relevant documents and allows limited negative feedback from only the top-ranked non-relevant document. For the experimental purpose, we consider the weights as $\alpha = 1$, $\beta = 1$, and $\gamma = 1$.

We also perform two different query reformulation based on local analysis. Generally, local analysis considers the top k most highly ranked documents for query expansion without any assistance from the user [12, 3]. However, in this work, we consider only the user selected relevant images for further analysis. At first, a simpler approach of query expansion is considered based on identifying most frequently occurring five keywords from user selected relevant documents. After selecting the additional keywords, the query vector is reformulated as $\mathbf{f}^m_{D_q}(Local1)$ by re-weighting its keywords based on the *tf-idf* weighting scheme and is re-submitted to the system as a new query. The other query reformulation approach is based on expanding the query with terms correlated to the query terms. Such correlated terms are those present in local clusters built from the relevant documents as indicated by the user. There are many ways to build a local cluster before performing any query expansion [12, 3]. For this work, a correlation matrix $C_{(|T_l| \times |T_l|)} = [c_{u,v}]$ is constructed [8] in which the rows and columns are associated with terms in a local vocabulary T_l. The element of this matrix $c_{u,v}$ is defined as

$$c_{u,v} = \frac{n_{u,v}}{n_u + n_v - n_{u,v}} \tag{4}$$

where, n_u is the number of local documents which contain term t_u, n_v is the number of local documents which contain term t_v, and $n_{u,v}$ is the number of local documents which contain both terms t_u and t_v. Here, $c_{u,v}$ measures the ratio between the number of local documents where both t_u and t_v appear and the total number of local documents where either t_u or t_v appear. If t_u and t_v have many co-occurrences in documents, then the value of $c_{u,v}$ increases, and the documents are considered to be more correlated. Now, given the correlation matrix C, we use it to build the local correlation cluster. For a query term $t_u \in D_q$, we consider the u-th row in C (i.e., the row with all the correlations for the keyword t_u). From that row, we return three largest correlation values $c_{u,l}, u \neq l$, and add corresponding terms t_l for query expansion. The process is continued for each query term and finally the query vector is reformulated as $\mathbf{f}^m_{D_q}(Local2)$ by re-weighting its keywords based on the *tf-idf* weighting scheme.

3 Content-based Image Retrieval Approach

In content-based image retrieval (CBIR), access to information is performed at a perceptual level based on automatically extracted low-level features (e.g., color, texture, shape, etc.) [13]. The performance of a content-based image retrieval (CBIR) system depends on the underlying image representation, usually in the form of a feature vector. To generate feature vectors, various global, semi-global, region-specific, and visual concept-based image features are extracted at different levels of abstraction. The MPEG-7 based Edge Histogram Descriptor (EHD) and Color Layout Descriptor (CLD) are extracted for image representation at global level [14]. To represent EHD as vector $\mathbf{f}^{\text{ehd}}$, a histogram with $16 \times 5 = 80$ bins is obtained. The CLD represents spatial layout of images in a very compact form in $YCbCr$ color space where Y is the luma component and Cb and Cr are the blue and red chroma components [14]. In this work, CLD with 10 Y, 3 Cb and 3 Cr coefficients is extracted to form a 16-dimensional feature vector $\mathbf{f}^{\text{cld}}$. The global distance measure between feature vectors of query image I_q and database image I_j is a weighted Euclidean distance measure and is defined as

$$\text{Dis}_{\text{global}}(I_q, I_j) = \omega_{\text{cld}} \text{Dis}_{\text{cld}}(\mathbf{f}^{\text{cld}}_{I_q}, \mathbf{f}^{\text{cld}}_{I_j}) + \omega_{\text{ehd}} \text{Dis}_{\text{ehd}}(\mathbf{f}^{\text{ehd}}_{I_q}, \mathbf{f}^{\text{ehd}}_{I_j}), \tag{5}$$

where, $\text{Dis}_{\text{cld}}(\mathbf{f}^{\text{cld}}_{I_q}, \mathbf{f}^{\text{cld}}_{I_j})$ and $\text{Dis}_{\text{ehd}}(\mathbf{f}^{\text{ehd}}_{I_q}, \mathbf{f}^{\text{ehd}}_{I_j})$ are the Euclidean distance measures for CLD and EHD respectively and ω_{cld} and ω_{ehd} are weights for each feature distance measure subject to

$0 \leq \omega_{\text{cld}}, \omega_{\text{ehd}} \leq 1$ and $\omega_{\text{cld}} + \omega_{\text{ehd}} = 1$ and initially adjusted with equal weights as $\omega_{\text{cld}} = 0.5$ and $\omega_{\text{ehd}} = 0.5$. For semi-global feature vector, a simple grid-based approach is used to divide the images into five overlapping sub-images [16]. Several moment based color and texture features are extracted from each of the sub-images and later they are combined to form a semi-global feature vector. The mean and standard deviation of each color channel in HSV color space are extracted form each overlapping sub-region of an image I_j. Various texture moment-based features (such as energy, maximum probability, entropy, contrast and inverse difference moment) are also extracted from the grey level co-occurrence matrix (GLCM) [15]. Color and texture feature vectors are normalized and combined to form a joint feature vector $\mathbf{f}^{\text{sg}}_{r_j}$ of each sub-image r and finally they are combined as the semi-global feature vector for an entire image as $\mathbf{f}^{\text{sg}}$. The semi-global distance measure between I_q and I_j is defined as

$$\text{Dis}_{\text{s-global}}(I_q, I_j) = \text{D}_{\text{sg}}(\mathbf{f}^{\text{sg}}_{I_q}, \mathbf{f}^{\text{sg}}_{I_j}) = \frac{1}{r} \sum_{r=1}^{5} \omega_r \text{Dis}_r(\mathbf{f}^{\text{sg}}_{r_q}, \mathbf{f}^{\text{sg}}_{r_j}) \tag{6}$$

where, $\text{Dis}_r(\mathbf{f}^{\text{sg}}_{r_q}, \mathbf{f}^{\text{sg}}_{r_j})$ is the Euclidean distance measure of the feature vector of region r and ω_r are the weights for the regions, which are set as equal initially.

Region-based image retrieval (RBIR) aims to overcome the limitations of global and semi-global retrieval approaches by fragmenting an image automatically into a set of homogeneous regions based on color and/or texture properties. Hence, we consider a local region specific feature extraction approach by fragmenting an image automatically into a set of homogeneous regions made up of (2 × 2) pixel blocks based on a fast k-means clustering technique. The image level distance between I_q and I_j is measured by integrating properties of all regions in the images. Suppose, there are M regions in image I_q and N regions in image I_j. Now, the image-level distance is defined as

$$\text{Dis}_{\text{local}}(I_q, I_j) = \frac{\sum_{i=1}^{M} w_{r_{i_q}} \text{Dis}_{r_{i_q}}(q, j) + \sum_{k=1}^{N} w_{r_{k_j}} \text{Dis}_{r_{k_j}}(j, q)}{2} \tag{7}$$

where $w_{r_{i_q}}$ and $w_{r_{k_j}}$ are the weights (e.g., number of image block as unit) for region r_{i_q} and region r_{k_j} of image I_q and I_j respectively. For each region $r_{i_q} \in I_q$, $\text{Dis}_{r_{i_q}}(q, j)$ is defined as the minimum Bhattacharyya distance [18] between this region and any region $r_{k_j} \in I_j$ as $\text{Dis}_{r_{i_q}}(q, j) = \min(\text{Dis}(r_{i_q}, r_{1_j}), \cdots, \text{Dis}(r_{i_q}, r_{N_j}))$. The Bhattacharyya distance is computed based on mean color vector and covariance matrix of color channels in HSV color space of each region. The details of the segmentation, local feature extraction and similarity matching schemes were described in our previous work in [16].

We also extract visual concept-based image features that is analogous to a keyword-based representation in text retrieval domain. The visual concepts depict perceptually distinguishable color or texture patches in local image regions. For example, a predominant yellow color patch can be presented either in an image of the sun or in a sunflower image. To generate a set of visual concepts analogous to a dictionary of keywords, we consider a fixed decomposition approach to generate a 16 × 16 grid based partition of images. Therefore, sample images from a training set are equally partitioned into 256 non-overlapping smaller blocks. To represent each block as a feature vector, color and texture moment-based features are extracted as described for the semi-global feature. To generate a coodbook of prototype concept vectors from the block features, we use a SOM-based clustering technique [17]. The basic structure of a SOM consists of two layers: an input layer and a competitive output layer. The input layer consists of a set of input node vector $X = \{\mathbf{x}_1, \cdots \mathbf{x}_i, \cdots \mathbf{x}_n\}$, $\mathbf{x}_i \in \Re^d$, while the output layer consists of a set of N neurons $C = \{c_1, \cdots c_j, \cdots c_N\}$, where each neuron c_j is associated with a weight vector $\mathbf{c}_j \in \Re^d$. After the weight vectors are determined through the learning process, each output neuron c_j resembles as a visual concept with the associated weight vector $\mathbf{c}_j$ as code vector of a codebook. To encode an image, it is also decomposed into an even gird-based partition, where the color and texture moment-based features are extracted from each block. Now, for joint color and texture moment-based feature vector of each block, the nearest output node $\mathbf{c}_k, 1 \leq k \leq N$ is identified by applying the Euclidean distance measure and the corresponding index k of the output node c_k is stored for that particular block of the image. Based on this encoding scheme, an image I_j can be represented as a vector $\mathbf{f}_{I_j}^{\mathrm{V-concept}} = [f_{1j}, \cdots, f_{ij}, \cdots f_{Nj}]^{\mathrm{T}}$, where each dimension corresponds to a concept index in the codebook. The element f_{ij} represents the frequency of occurrences of c_i appearing in I_j. For this work, codebooks of size of 400 (e.g.,20×20 units) are constructed for the photographic and medical collection by manually selecting 2% images from each collection as training set. Since, the concept-based feature space is closely related to the keyword-based feature space of documents, we apply the cosine measure to compare image I_q and I_j as described in equation (1).

3.1 Visual Query Refinement by Relevance Feedback

This section presents the visual query refinement approach at different levels of image representation. The query refinement is closely related to the approach in [9]. It is assumed that, all positive feedback images at some particular iteration belong to user perceived visual and/or semantic category and obey the Gaussian distribution to form a cluster in the feature space. We consider the rest of the images as irrelevant and they may belong to different semantic categories. However, we do not consider the irrelevant images for query refinement. The modified query vector at a particular iteration is represented as the mean of the relevant image vectors

$$\mathbf{f}_{I_q}^{x^m} = \frac{1}{|R|} \sum_{\mathbf{f}_{I_l} \in R} \mathbf{f}_{I_l}^{x} \tag{8}$$

where, R is the set of relevant image vectors and $x \in \{\mathrm{global}, \mathrm{sg}, \mathrm{V-concept}\}$. Next, the covariance matrix of the positive feature vectors is estimated as

$$\mathbf{C}^x = \frac{1}{|R|-1} \sum_{l=1}^{|R|} (\mathbf{f}_{I_l}^{x^m} - \mathbf{f}_{I_q}^{x})(\mathbf{f}_{I_l}^{x^m} - \mathbf{f}_{I_q}^{x})^T \tag{9}$$

However, singularity issue will arise in covariance matrix estimation if fewer training samples or positive images are available compared to the feature dimension (as will be the case in user feedback images). So, we add regularization to avoid singularity in matrices as follows[19]:

$$\hat{\mathbf{C}^x} = \alpha \mathbf{C}^x + (1-\alpha)\mathbf{I} \tag{10}$$

for some $0 \leq \alpha \leq 1$ and $\mathbf{I}$ is the identity matrix. After generating the mean vector and covariance matrix for a feature $x \in \{\mathrm{global}, \mathrm{sg}, \mathrm{V-concept}\}$, we adaptively adjust the distance measure functions in equations (5) and (6) with the following Mahalanobis distance measures [18] for query image I_q and database image I_j as

$$\mathrm{Dis}_{\mathrm{x}}(I_q, I_j) = (\mathbf{f}_{I_q}^{x^m} - \mathbf{f}_{I_j}^{x})^T \hat{\mathbf{C}^x}^{-1} (\mathbf{f}_{I_q}^{x^m} - \mathbf{f}_{I_j}^{x}) \tag{11}$$

The Mahalanobis distance differs from the Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant, i.e. not dependent on the scale of measurements [18]. We did not perform any query refinement for region-specific feature due to its variable feature dimension for variable number of regions in each image.

4 Combination of Evidence by Dynamic Weight Update

In recent years, the category of work known as data fusion or multiple-evidence described a range of techniques in information retrieval whereby multiple pieces of information are combined to achieve improvements in retrieval effectiveness [10, 11]. These pieces of information can take many forms including different query representations, different document (image) representations, and different retrieval strategies used to obtain a measure of relationship between a query and a document (image). Motivated by this paradigm, in Sections 2 and 3, we described multiple textual query and image representation schemes. This section presents an adaptive linear combination approach based on relevance feedback information. One of the most commonly used approaches in data fusion is the linear combination of similarity scores. For our multi-modal retrieval purpose, let us consider q as a multi-modal query which has an image part as I_q and a document part as annotation file as D_q. In a linear combination scheme, the similarity between q and a multi-modal item j, which also has two parts (e.g., image I_j and text D_j), is defined as

$$\mathrm{Sim}(q,j) = \omega_I \mathrm{Sim}_{\mathrm{I}}(I_q, I_j) + \omega_D \mathrm{Sim}_{\mathrm{D}}(D_q, D_j) \tag{12}$$

where ω_I and ω_D are inter-modality weights within the text or image feature space, which subject to $0 \leq \omega_I, \omega_D \leq 1$ and $\omega_I + \omega_D = 1$. Now, the image based similarity is again defined as the linear combination of similarity measures in different level of image representation as

$$\mathrm{Sim}_{\mathrm{I}}(I_q, I_j) = \sum_{IF} \omega_I^{IF}\ \mathrm{Sim}_I^{IF}(I_q, I_j) \tag{13}$$

where $IF \in \{\mathrm{global}, \mathrm{semi-global}, \mathrm{region}, \mathrm{V-concept}\}$ and ω^{IF} are the weights within the different image representation schemes (e.g., intra-modality weights). On the other hand, the text based similarity is defined as the linear combination of similarity matching based on different query representation schemes.

$$\mathrm{Sim}_{\mathrm{D}}(D_q, D_j) = \sum_{QF} \omega_D^{QF}\ \mathrm{Sim}_D^{QF}(D_q, D_j) \tag{14}$$

where $QF \in \{\mathrm{Rocchio}, \mathrm{Ide}, \mathrm{Local1}, \mathrm{Local2}\}$ and ω^{QF} are the weights within the different query representation schemes.

The effectiveness of the linear combination depends mainly on the choice of the different inter and intra-modality weights. We use a dynamic weight updating method in linear combination schemes by considering both precision and rank order information of top retrieved K images. Before any fusion, the distance scores of each representation are normalized and converted to the similarity scores with a range of $[0,1]$ as $\mathrm{Sim}(q,j) = 1 - \frac{\mathrm{Dis}(q,j) - \min(\mathrm{Dis}(q,j))}{\max(\mathrm{Dis}(q,j)) - \min(\mathrm{Dis}(q,j))}$, where $\min(\cdot)$ and $\max(\cdot)$ are the minimum and maximum distance scores. In this approach, an equal emphasis is given based on their weights to all the features along with their similarity matching functions initially. However, the weights are updated dynamically during the subsequent iterations by incorporating the feedback information from the previous round. To update the inter-modality weights (e.g., ω_I and ω_D), we at first need to perform the multi-modal similarity matching based on equation (12). After the initial retrieval result with a linear combination of equal weights (e.g., $\omega_I = 0.5$ and $\omega_D = 0.5$), a user needs to provide a feedback about the relevant images from the top K returned images. For each ranked list based on individual similarity matching, we also consider top K images and measure the effectiveness of a query/image feature as

$$\mathrm{E}(D \text{ or I}) = \frac{\sum_{\mathrm{i=1}}^{\mathrm{K}} \mathrm{Rank(i)}}{\mathrm{K}/2} * \mathrm{P(K)} \tag{15}$$

where $\mathrm{Rank(i)} = 0$ if image in the rank position i is not relevant based on user's feedback and $\mathrm{Rank(i)} = (\mathrm{K} - \mathrm{i})/(\mathrm{K} - 1)$ for the relevant images. Here, $P(K) = R_K/K$ is the precision at top K, where R_k be the number of relevant images in the top K retrieved result. Hence, the equation (15) is basically the product of two factors, rank order and precision. The raw performance scores obtained by the above procedure are then normalized by the total score as $\hat{E}(D) = \hat{\omega}_D = \frac{E(D)}{E(D)+E(I)}$ and $\hat{E}(I) = \hat{\omega}_I = \frac{E(I)}{E(D)+E(I)}$ to generate the updated text and image feature weights respectively. For the next iteration of retrieval with the same query, these modified weights are utilized for the multi-modal similarity matching function as

$$\mathrm{Sim}(q,j) = \hat{\omega}_I \mathrm{Sim}_{\mathrm{I}}(I_q, I_j) + \hat{\omega}_D \mathrm{Sim}_{\mathrm{D}}(D_q, D_j) \tag{16}$$

This weight updating process might be continued as long as users provide relevant feedback information or until no changes are noticed due to the system convergence.

In a similar fashion, to update the intra-modality weights (e.g., ω_D^{QF} and ω_I^{IF}), we need to consider the top K images in individual result list. So, for image-based similarity in equation (13), we consider the result lists of different image features of $IF \in \{\mathrm{global}, \mathrm{semi-global}, \mathrm{region}, \mathrm{V-concept}\}$ and measure their weights by using equation (15) for the next retrieval iteration. On the other hand, for text-based similarity in equation (14), the top K images in result lists of different query features of $QF \in \{\mathrm{Rocchio}, \mathrm{Ide}, \mathrm{Local1}, \mathrm{Local2}\}$ are considered and text-level weights are determined in a similar way by applying equation (15).

5 Sequential approach with pre-filtering and re-ordering

This section describes the process about how to interact with both the modalities in a user's perceived semantical and sequential way. Since a query can be represented with both keywords and visual features, it can be initiated either by the keyword-based search or by the visual example image search. However, we consider a pre-filtering and re-ranking approach based on the image search in the filtered image set which is obtained previously by the textual search. It would be more appropriate to perform a text-based search at first due to the higher level information content and latter use visual only search to refine or re-rank the top returned images by the textual search. In this method, combining the results of the text and image based retrieval is a matter of re-ranking or re-ordering of the images in a text-based pre-filtered result set. The steps involved in this approach are as follows:

Step 1: Initially, for a multi-modal query q with a document part as D_q, perform a textual search with vector $\mathbf{f}_{D_q}$ and rank the images based on the ranking of the associated annotation files by applying equation (1).

Step 2: Obtain user's feedback from top retrieved $K = 30$ images about relevant and irrelevant images for the textual query refinement.

Step 3: Calculate the optimal textual query vectors as $\mathbf{f}^m_{D_q}(Rocchio), \mathbf{f}^m_{D_q}(Ide), \mathbf{f}^m_{D_q}(Local1)$ and $\mathbf{f}^m_{D_q}(Local2)$.

Step 4: Re-submit the modified query vectors in the text engine and merge the results with an equal weighting in similarity matching in equation (14).

Step 5: Continue steps 2 to 4 with dynamically updated weights based on equation (15) until the user switch to visual only search.

Step 6: Extract different features as $\mathbf{f}^{\mathrm{global}}_{I_q}, \mathbf{f}^{\mathrm{sg}}_{I_q}, \mathbf{f}^{\mathrm{local}}_{I_q}$, and $\mathbf{f}^{\mathrm{V-concept}}_{I_q}$ for the multi-modal query q with an image part as I_q.

Step 7: Perform visual only search in top $L = 1000$ images retrieved by text-based search and rank them based on the similarity values by applying equation (13) with equal feature weighting.

Step 8: Obtain user's feedback from top retrieved $K = 30$ images about the relevant images and perform visual query refinement as $\mathbf{f}^{x^m}_{I_q}$, where $x \in \{\mathrm{global}, \mathrm{sg}, \mathrm{V-concept}\}$ at a particular iteration.

Step 9: At next iteration, calculate the feature weights based on equation (15) and apply it to equation (13) for ranked based retrieval result.

Step 10: Continue steps 8 and 9, until the user is satisfied or the system converges.

The process flow diagram of the sequential search approach is shown in Figure 1. For this approach, the text-based search with query reformulation is performed first as shown in the (1) left portion of the figure and image-based search is performed in the filtered image set as shown in the (1) right portion of the figure 1.

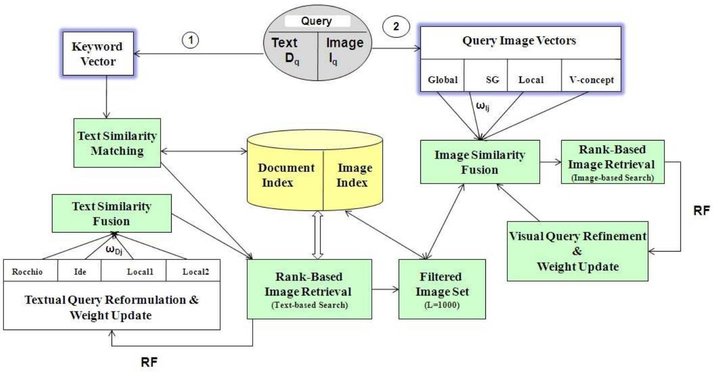

Figure 1: Process flow diagram of the sequential approach

6 Simultaneous approach with linear combination

This section describes our approach of simultaneous multi-modal search. Here, textual and content-based search are performed simultaneously from the beginning and the results are combined with an adaptive linear combination scheme as described in Section4. The steps involved in this approach are as follows:

Step 1: Initially, for a multi-modal query q with a document part as D_q and an image part as I_q, extract textual query vector as $\mathbf{f}_{D_q}$ and different image feature vectors as $\mathbf{f}_{I_q}^{\text{global}}, \mathbf{f}_{I_q}^{\text{sg}}, \mathbf{f}_{I_q}^{\text{local}}$, and $\mathbf{f}_{I_q}^{\text{V}-\text{concept}}$.

Step 2: Perform a multi-modal search to rank the images based on equation (12), where $\text{Sim}_{\text{D}}(D_q, D_j)$ is initially performed through $\text{Sim}_{\text{text}}(D_q, D_j)$ equation (1) and $\text{Sim}_{\text{I}}(I_q, I_j)$ is performed through equation (13) with initially equal weighting in both inter and intra-modality weights.

Step 3: Obtain user's feedback from top retrieved $K = 30$ images about relevant and irrelevant images for both textual and visual query refinement and for dynamically update the weights.

Step 4: Based on the feedback information, calculate the optimal textual query vectors as $\mathbf{f}_{D_q}^{m}(Rocchio), \mathbf{f}_{D_q}^{m}(Ide), \mathbf{f}_{D_q}^{m}(Local1)$ and $\mathbf{f}_{D_q}^{m}(Local2)$ and image query vectors as $\mathbf{f}_{I_q}^{x^m}$, where $x \in \{\text{global}, \text{sg}, \text{V}-\text{concept}\}$ and update the inter and intra-modality weights based on equation (15).

Step 5: Re-submit the modified textual and image query vectors to the system and apply multi-modal similarity matching based on equation (16), where $\text{Sim}_{\text{D}}(D_q, D_j)$ is performed through equation (14) and $\text{Sim}_{\text{I}}(I_q, I_j)$ is performed through equation (13).

Step 6: Continue steps 3 to 5, until the user is satisfied or the system converges.

The process flow diagram of the above multi-modal simultaneous search approach is shown in Figure 2. For this approach, both text and image-based search are performed simultaneously as shown in left and right portions of Figure 2.

6.0.1 Analysis of the submitted runs

The types and performances of the different runs are shown in Table 1 and Table 2 for the ad-hoc retrieval of the photographic and medical collections respectively. In all these runs, only English is used as the source and target language without any translation for the text-based retrieval approach. We submitted five different runs for the ad-hoc retrieval of the photographic collection, where first two runs are based on text only search and last three runs are based on mixed modality search as shown in Table 1. For the first run *"CINDI-TXT-ENG-PHOTO"*, we performed only a manual text-based search without any query expansion as our base run. This run achieved a MAP

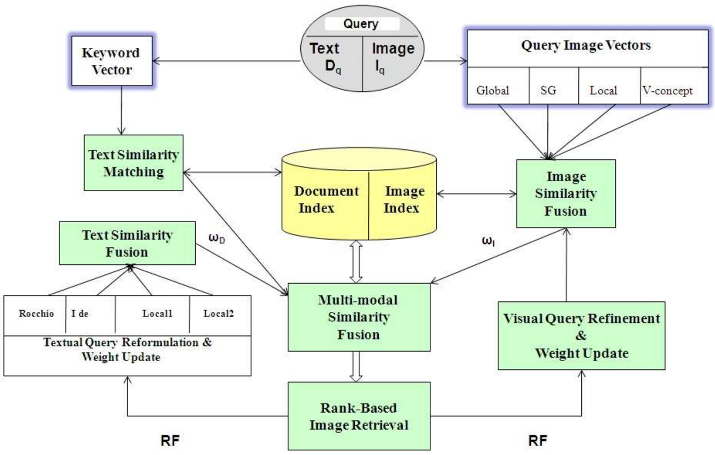

Figure 2: Process flow diagram of the simultaneous approach

Table 1: Results of the ImageCLEFphoto Retrieval task

Run ID	Modality	Run Type	QE/RF	MAP	BPREF
CINDI-TXT-ENG-PHOTO	TXT	Manual	NOFB	0.1529	0.1426
CINDI-TXT-QE-PHOTO	TXT	Manual	FBQE	0.2637	0.2515
CINDI-TXT-QE-IMG-RF-RERANK	MIXED	Manual	FBQE	0.2336	0.2398
CINDI-TXTIMG-FUSION-PHOTO	MIXED	Manual	NOFB	0.1483	0.1620
CINDI-TXTIMG-RF-PHOTO	MIXED	Manual	FBQE	0.1363	0.1576

score of 0.1529 and ranked 140th out of 476 submissions (e.g., within the top 30%). Our second run *"CINDI-TXT-QE-PHOTO"* achieved the best MAP score (0.2637) among all our submitted runs and ranked 21st for this year competition. In this run, we performed two iterations of manual feedback for textual query expansion and combination based on dynamic weight update schemes for text only retrieval as described in Sections 2 and 4. The rest of the runs are based on multi-modal approach, where in the third run *"CINDI-TXT-QE-IMG-RF-RERANK"*, we performed the sequential approach with pre-filtering and re-ordering as described in subsection 5 with two iterations of manual feedback in both text and image-based searches. However, the re-ordering approach did not improve the result as a whole (e.g., ranked 32nd) in terms of MAP score (0.2336) as compared to the only textual query expansion approach of our best run. The main reason might be due to the fact that the majority of the query topics are more semantically oriented, where visual search is not suitable or feasible at all. However, this run might perform well where queries have both textual and distinct visual properties, such as query topic number 15 as *"night shots of cathedrals"* or query topic number 24 as *"snowcapped building in Europe"*. For the fourth run *"CINDI-TXTIMG-FUSION-PHOTO"*, we performed a simultaneous retrieval approach without any feedback information with a linear combination of weights as $\omega_D = 0.7$ and $\omega_I = 0.3$ and for the fifth run *"CINDI-TXTIMG-RF-PHOTO"*, two iterations of manual relevance feedback are performed as described in Section 6. However, these two runs did not perform well in terms of MAP score as compared to the sequential approach due to early combination and nature of the queries as described earlier.

For the image retrieval task in the medical collections, we submitted seven runs this year. However, due to few errors (such as duplicate entry and reference image as 0.jpg in the result set), three of our runs could not produce performance report by evaluating with the *trec-eval* program. This is mainly due to reason of directly using reference images from the annotation

Table 2: Results of the Medical Retrieval task

Run ID	Modality	Run Type	QE/RF	MAP	R-prec
CINDI-IMG-FUSION	IMAGE	Manual	NOFB	0.0333	0.0532
CINDI-IMG-FUSION-RF	IMAGE	Manual	FBQE	0.0372	0.0549
CINDI-TXT-IMAGE-LINEAR	MIXED	Manual	NOFB	0.1659	0.2196
CINDI-TXT-IMG-RF-LINEAR	MIXED	Manual	FBQE	0.0823	0.1168

files instead of using the link XML file as provided. We are currently fixing this problem and later analyze and report the results of these runs. Table 2 shows the official result of the four runs out of our seven submitted runs. In the first run *"INDI-IMG-FUSION"*, we performed only a visual only search based on various image feature representation schemes as described in Section 3 without any feedback information and with a linear combination of equal feature weights. For the second run *"CINDI-IMG-FUSION-RF"*, we performed only one iteration of manual feedback for visual query refinement and combined the similarity matching functions based on the dynamic weight updating scheme. For this run we achieved a MAP score of 0.0372, which is slightly better then the score (0.0333) achieved by the first run without any relevance feedback information. However, compared to the the text-based approaches the performances are very low as it happened in previous years of ImageCLEFmed. For the third run *"CINDI-TXT-IMAGE-LINEAR"*, we performed a simultaneous retrieval approach without any feedback information with a linear combination of weights as $\omega_D = 0.7$ and $\omega_I = 0.3$ and for the fourth run *"CINDI-TXT-IMG-RF-LINEAR"*, two iterations of manual relevance feedback are performed similar to the last two runs of photographic retrieval task. From Table 2, it is clear that combining both modalities for the medical retrieval task is far better then using only a single modality (e.g., only image) and we achieved the best MAP score as 0.1483 among all our submissions for this task.

7 Conclusion

This paper presents the ad-hoc image retrieval approaches of CINDI research group for ImageCLEF 2007. We submitted several runs with different combination of methods, features and parameters. We investigated with cross-modal interaction and fusion approaches for the retrieval of the photographic and medical image collections. The description of the runs and analysis of the results are discussed in this paper.

References

[1] M. Grubinger, P. Clough, A. Hanbury, and H. Müller, Overview of the ImageCLEF 2007 Photographic Retrieval Task, *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, Sep. 2007.

[2] H. Müller, T. Deselaers, E. Kim, C. Kalpathy, D. Jayashree, M. Thomas, P. Clough, W. Hersh, Overview of the ImageCLEFmed 2007 Medical Retrieval and Annotation Tasks, *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, Sep. 2007.

[3] R. Baeza-Yates and B. Ribiero-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

[4] G. Salton and C. Buckley, Improving retrieval performance by relevance feedback, *Journal of the American Society for Information Science*, vol. 41(4), pp. 288–297, 1990.

[5] Y. Rui, T. S. Huang, Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval, *IEEE Circuits Syst. Video Technol.*, vol. 8, pp. 644–655, 1999.

[6] J.J. Rocchio, Relevance feedback in information retrieval. In *The SMART Retrieval System - Experiments in Automatic Document Processing*, pp. 313–323, Englewood Cliffs, NJ, Prentice Hall, Inc. 1971.

[7] E. Ide, New experiments in relevance feedback, In *The SMART retrieval system - Experiments in Automatic Document Processing*, pp 337–354. 1971.

[8] Y. Ogawa, T. Morita, and K. Kobayashi, A fuzzy document retrieval system using the keyword connection matrix and a learning method, *Fuzzy Sets and Systems*, vol. 39 pp. 163–179, 1991.

[9] Y. Ishikawa, R. Subramanya and C. Faloutsos, MindReader: Querying Databases Through Multiple Examples, *24th Internat. Conf. on Very Large Databases*, New York, pp. 24–27, 1998.

[10] E.A. Fox and J.A. Shaw, Combination of Multiple Searches, *Proc. of the 2nd Text Retrieval Conference (TREC-2)*, NIST Special Publication 500-215, pp. 243-252, 1994.

[11] J.H. Lee, Combining Multiple Evidence from Different Properties of Weighting Schemes, *Proc. of the 18th Annual ACM-SIGIR*, pp. 180–188, 1995.

[12] R. Attar and A.S. Fraenkel, Local feedback in full-text retrieval systems, *Journal of ACM*, vol. 24 (3), pp. 397–417, 1977.

[13] A. Smeulder, M. Worring, S. Santini, A. Gupta, R. Jain, Content-Based Image Retrieval at the End of the Early Years, *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 22, pp. 1349–1380, 2000.

[14] B.S. Manjunath, P. Salembier, T. Sikora, (eds.), *Introduction to MPEG-7- Multimedia Content Description Interface*, John Wiley Sons Ltd. pp. 187–212, 2002.

[15] R.M. Haralick, Shanmugam, and I. Dinstein, Textural features for image classification, *IEEE Trans System, Man, Cybernetics*, vo;. 3, pp. 610–621, 1973.

[16] M.M. Rahman, B.C. Desai, and P. Bhattacharya, A Feature Level Fusion in Similarity Matching to Content-Based Image Retrieval, *Proc. 9th Internat Conf Information Fusion*, 2006.

[17] T. Kohonen, *Self-Organizing Maps*, Springer-Verlag, Heidelberg. 2nd ed. 1997.

[18] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. Academic Press, 1990.

[19] J. Friedman, Regularized Discriminant Analysis, *Journal of American Statistical Association*, vol. 84, pp. 165–175, 2002.