

CINDI at ImageCLEF 2006: Image Retrieval & Annotation Tasks for the General Photographic and Medical Image Collections

M. M. Rahman, Varun Sood, 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 our techniques used and their analysis for the runs made and the results submitted by the CINDI group for the task of the image retrieval and automatic annotation of ImageCLEF 2006. For the ad-hoc image retrieval from both the photographic and medical image collections, we have experimented with cross-modal (image and text) interaction and integration approaches based on the relevance feedback in the form of textual query expansion and visual query point movement with adaptive similarity matching functions. Experimental results show that our approaches performed well compared to initial visual or textual only retrieval without any user interactions or feedbacks. We are ranked first and second and achieved the highest MAP score (0.3850) for the ad-hoc retrieval in the photographic collection (IAPR) among all the submissions. For the automatic annotation tasks for both the medical (IRMA) and object collections (LTU), we have experimented with a classifier combination approach, where several probabilistic multi-class SVM classifiers with features at different levels as inputs are fused with several combination rules to predict the final probability score of each category as image annotation. Analysis of the results of the different runs we made for both the image retrieval and annotation tasks 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, Machine learning, Performance, Experimentation

Keywords

Content-based image retrieval, Vector space model, Feature extraction, Query expansion, Relevance feedback, Classification, Support vector machine

1 Introduction

For the 2006 ImageCLEF workshop, CINDI research group has participated in four different tasks of ImageCLEF track: an ad-hoc retrieval from a photographic collection, ad-hoc retrieval from a medical collection, and automatic annotation of the medical and the object data sets [1, 2]. This paper presents the methodologies, results and analysis of the runs of each of the tasks separately.

2 Ad-hoc retrieval from photographic collection

Our main goal of the ad-hoc retrieval task is to investigate the effectiveness of combining text and image by involving user in the retrieval loop in the form of relevance feedback. We have experimented with a cross-modal approach of image retrieval which integrates visual information based on purely low-level image content and semantical information from the associated annotated text files. The advantages of both modalities are exploited by involving the users in the retrieval loop for the cross-modal interaction and integration in similarity matching.

For the text-based retrieval, the keywords from the annotated files are extracted and indexed with the help of the vector space model paradigm [3]. In order to perform the query expansion on the textual search, additional keywords are extracted for the query based on the positive feedbacks from the user. For the content-based search, a query point movement and an adjustment of the similarity matching functions are performed based on the estimation of the mean and covariance matrix from the feature vectors of the positive feedback images. Finally, a ranked-based ordered list of images is obtained by a pre-filtering approach which integrates the scores from both the text and image search-based result lists.

2.1 Text retrieval approach

For the keyword-based search on the annotated text files, we have utilized a simple but effective information retrieval (IR) tool by Raymond Mooney at the Texas University [4]. However, we have performed several modifications to the original library according to the experimental requirements, such as allowing recursive indexation of the text files that are stored in directories, expanding the stop word list by adding several common words specific to the experimental domain and modifying the term weighting scheme with the query expansion. For the text-based indexing, keywords are extracted from all the associated annotation files by ignoring all the tags as stop words.

The indexing technique is based on the popular vector space model (VSM) of IR [3]. In this model, texts and queries are represented as vectors in a N -dimensional space, where N is the number of keywords in the collection. So, each document j can be represented as a vector as:

$$\mathbf{D}_j = \langle w_{1j}, \dots, w_{Nj} \rangle \quad (1)$$

The element w_{ij} represents the weights of the keyword w_i appearing in document j and can be weighted in a variety of ways. One common scheme is *term-frequency-inverse document frequency* (TF-IDF) weighting. Both global weight and local weight are considered in this approach [3]. A global weight indicates the overall importance of that component in the feature vector across the whole image collection. A local weight is applied to each element indicating the relative importance of the component within its vector. The local weight is denoted as $L_{i,j} = \log(f_{i,j}) + 1$, where $f_{i,j}$ is the frequency of occurrence of keyword w_i in document j . The global weight is the inverse document frequency and denoted by G_i where $G_i = \log(M/M_i) + 1$, for $i = (1, \dots, N)$, where M_i be the number of documents in which w_i is found and M is the total number of documents in the collection. Finally, the element w_{ij} is expressed as the product of local and global weight: hence $w_{ij} = L_{i,j} * G_i$ [3].

The vector space model is based on the assumption that similar documents will be represented by similar vectors in the N -dimensional vector space. In particular, similar documents are expected to have small angles between their corresponding vectors. Hence, the cosine similarity measure is

adopted between feature vectors of the query document q and database document j as follows [3]:

$$S_{\text{text}}(q, j) = S_{\text{text}}(\mathbf{D}_q, \mathbf{D}_j) = \frac{\sum_{i=1}^N w_{iq} * w_{ij}}{\sqrt{\sum_{i=1}^N (w_{iq})^2} * \sqrt{\sum_{i=1}^N (w_{ij})^2}} \quad (2)$$

where, \mathbf{D}_q and \mathbf{D}_j are the query and document vector respectively. The advantage of the VSM includes a ranked result of the retrieved documents (as well as the associated images) which would be useful when we fuse the results from both the keyword and content-based image retrieval.

2.2 Content-based image retrieval approach

The performance of a Content-based image retrieval approach (CBIR) system depends on the underlying image representation, usually in the form of a feature vector [5]. Based on the previous experiments [6], we have found that the image features at different levels are complementary in nature and together they could contribute to effectively distinguish the images of different semantic categories. Hence, to generate the feature vectors, we have extracted the low-level global, semi-global and region specific local features for the image representation at different levels of abstraction.

2.2.1 Feature extraction and similarity matching

In this work, the MPEG-7 based Edge Histogram Descriptor (EHD) and Color Layout Descriptor (CLD) are extracted for image representation at the global level [7]. To represent the global shape feature, the spatial distribution of edges are utilized by the EHD descriptor. A histogram with $16 \times 5 = 80$ bins is obtained, corresponding to a feature vector \mathbf{f}^{EHD} , having a dimension of 80 [7]. The CLD represents the spatial layout of the images in a very compact form [7]. It is obtained by applying the discrete cosine transform (DCT) on the 2-D array of local representative colors in $YCbCr$ color space. In this work, CLD with 10 Y , 3 Cb and 3 Cr coefficients is extracted to form a 16-dimensional feature vector \mathbf{f}^{CLD} .

Now, for comparing the query image Q and the target image T in the database based on the global features, a weighted Euclidean distance measure is utilized as

$$\text{DIS}_{\text{global}}(Q, T) = \omega_{\text{CLD}} D_{\text{CLD}}(Q, T) + \omega_{\text{EHD}} D_{\text{EHD}}(Q, T), \quad (3)$$

where, $D_{\text{CLD}}(Q, T) = \|\mathbf{f}_Q^{\text{CLD}} - \mathbf{f}_T^{\text{CLD}}\|^2$ and $D_{\text{EHD}}(Q, T) = \|\mathbf{f}_Q^{\text{EHD}} - \mathbf{f}_T^{\text{EHD}}\|^2$ are the Euclidean distance measures for CLD and EHD feature vector respectively and ω_{CLD} and ω_{EHD} are weights for each feature distance measure subject to $\omega_{\text{CLD}} + \omega_{\text{EHD}} = 1$ and adjusted as $\omega_{\text{CLD}} = 0.4$ and $\omega_{\text{EHD}} = 0.6$ in the experiment.

For semi-global feature vector, a simple grid-based approach is used to divide the images into five overlapping sub-images [6]. 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. For moment-based color feature, the first (mean) and second (standard deviation) central moments of each color channel in HSV color space are extracted. Texture features are extracted from the grey level co-occurrence matrix (GLCM) [8]. GLCM is defined as a sample of the joint probability density of the gray levels of two pixels separated by a given displacement. Second order moments, such as energy, maximum probability, entropy, contrast and inverse difference moment are measured based on the GLCM. Color and texture feature vectors are normalized and combined to form a joint feature vector of 11-dimensions (6 for color and 5 for texture) for each sub-region to finally generate a 55-dimensional (5×11) semi-global feature vector \mathbf{f}^{SG} . For the semi-global distance measure between Q and T , we also utilized the Euclidean distance measure as

$$\text{DIS}_{\text{semi-global}}(Q, T) = \|\mathbf{f}_Q^{\text{SG}} - \mathbf{f}_T^{\text{SG}}\|^2 \quad (4)$$

We have also considered a local region specific feature extraction approach by fragmenting an image automatically into a set of homogeneous regions based on a fast k-means clustering technique. To

represent each region with local features, we consider information on weight (i.e, number of pixels) and color-texture as in [6]. Color feature $\mathbf{f}_{R_i}^c$ of each region R_i is a 3-D vector and is represented by the K-means cluster center, i.e., the average value for each of the three color channels in HSV space of all the image pixels in this region. Texture feature of each region is measured in an indirect way by considering the cross-correlation among color channels due to the off diagonal of the 3×3 covariance matrix of the region R_i .

To compute the region specific distance measure between two regions R_i and R_j of Q and T respectively, we apply the Bhattacharyya distance metric [9] as follows:

$$D(R_i, R_j) = \frac{1}{8}(\mathbf{f}_{R_i}^c - \mathbf{f}_{R_j}^c)^T \left[\frac{(C_{Q_{R_i}} + C_{T_{R_j}})}{2} \right]^{-1} (\mathbf{f}_{R_i}^c - \mathbf{f}_{R_j}^c) + \frac{1}{2} \ln \frac{\left| \frac{(C_{Q_{R_i}} + C_{T_{R_j}})}{2} \right|}{\sqrt{|C_{Q_{R_i}}| |C_{T_{R_j}}|}} \quad (5)$$

where $\mathbf{f}_{R_i}^c$ and $\mathbf{f}_{R_j}^c$ are the region feature vectors, and $C_{Q_{R_i}}$ and $C_{T_{R_i}}$ are the covariance matrices of region R_i and R_j of query image Q and target image T respectively. Finally, the image-level distance between Q and T is measured as

$$\text{DIS}_{\text{local}}(Q, T) = \frac{\sum_{i=1}^M w_{Q_{R_i}} R_i(T) + \sum_{j=1}^N w_{T_{R_j}} R_j(Q)}{2} \quad (6)$$

where $w_{Q_{R_i}}$ and $w_{T_{R_j}}$ are the weights for region i of image Q and region j of image T respectively. For each region $i \in M$ in Q , $R_i(T)$ is defined as the minimum distance between this region and any region $j \in N$ in image T and in a similar way $R_j(Q)$ is computed.

The overall image level similarity is measured by fusing of a weighted combination of individual similarity measures. Once the distance functions are measured as above, they are normalized and converted to similarity measure, which in general is the converse of a distance function. After the similarity measures of each representation are determined as $S_{\text{global}}(Q, T)$, $S_{\text{semi-global}}(Q, T)$, and $S_{\text{local}}(Q, T)$, we aggregate or fuse them into a single similarity matching function as follows:

$$S_{\text{image}}(Q, T) = w_g S_{\text{global}}(Q, T) + w_{\text{sg}} S_{\text{semi-global}}(Q, T) + w_l S_{\text{local}}(Q, T) \quad (7)$$

Here, w_g , w_{sg} and w_l are non-negative weighting factors of different feature level similarities with normalization $w_g + w_{\text{sg}} + w_l = 1$. For the retrieval experiments, they are selected as $w_g = 0.4$, $w_{\text{sg}} = 0.3$ and $w_l = 0.3$.

2.3 Cross-modal interaction with relevance feedback

In a practical multi-modal image retrieval system, the user at first might want to search images with keywords as it is more convenient and semantically more appropriate. However, a short query (e.g., query topic) with few keywords might not be enough to incorporate the user perceived semantics to the retrieval system. Hence, a query expansion process is required to add additional keywords and modify the weight of the keywords in the original query vector. In this paper, a simpler approach of query expansion is considered based on identifying useful terms or keywords from the associated annotation files for the images.

The approach of the textual query expansion based on the relevance feedback (RF) is as follows: the user provides the initial query topic and the system extracts from it a set of keywords as the initial textual query vector $\mathbf{D}_{q(0)}$. This query vector is used to retrieve K most similar images from associated text documents based on the cosine similarity measure described in section 2.1. If the user is not satisfied with the result, then the system will allow the user to select a set of relevant or positive images close to the semantics of the initial textual query topic. Next, the system will extract all the keywords from the annotation files associated with the positive feedback images.

After extracting the additional keywords, the query vector will be adjusted as $\mathbf{D}_{q(i)}$ at iteration i by re-weighting its keywords by following the TF-IDF and re-submitted to the system as the query for the next iteration. This process may continue for several iterations until the user is satisfied with the result.

However, since we have a multi-modal system, it will not be wise to perform query expansion by just using one particular modality (e.g., only text). Visual features of images also play an important part in distinguishing different semantical/visual categories. Therefore, we also need to perform RF with content-based image search for better precision [11]. In this scenario, like textual query expansion, user might provide the initial image query vector $\mathbf{f}_{Q(0)}$ to retrieve K most similar images based on the similarity measure function in equation (7). In the next iteration (either from the textual or image-based feedback), user might select a set of relevant images compared to the initial query image. It is assumed that, all the positive feedback images $Pos(\mathbf{f}_{Q(i)})$ at some particular iteration i will belong to the user perceived semantic category and obey the Gaussian distribution to form a cluster in the feature space.

Let, N_{Pos} be the number of positive feedback images at iteration i and $\mathbf{f}_{T_j} \in \mathfrak{R}^d$ is be the feature vector that represents j -th image for $j \in \{1, \dots, N_{Pos}\}$, then the new query point at iteration i is estimated as $\mathbf{f}_{Q(i)} = \frac{1}{N_{Pos}} \sum_{j=1}^{N_{Pos}} \mathbf{f}_{T_j}$ as the mean vector of positive images and covariance matrix is estimated as $\mathbf{C}_{(i)} = \frac{1}{N_{Pos}-1} \sum_{j=1}^{N_{Pos}} (\mathbf{f}_{T_j} - \mathbf{f}_{Q(i)})(\mathbf{f}_{T_j} - \mathbf{f}_{Q(i)})^T$. However, singularity issue will arise in covariance matrix estimation if fewer than $d + 1$ training samples or positive images are available as will be the case in user feedback images. So, we add regularization to avoid singularity in matrices as follows[12]:

$$\hat{\mathbf{C}}_{(i)} = \alpha \mathbf{C}_{(i)} + (1 - \alpha) \mathbf{I} \quad (8)$$

for some $0 \leq \alpha \leq 1$ and \mathbf{I} is the $d \times d$ identity matrix.

After generating the mean vector and covariance matrix of the positive images, we adaptly adjust the Euclidean distance measures of various feature representation with the following Mahalanobis distance measure [9]:

$$\text{DIS}_{\text{Maha}}(Q, T) = (\mathbf{f}_{Q(i)} - \mathbf{f}_T)^T \hat{\mathbf{C}}_{(i)}^{-1} (\mathbf{f}_{Q(i)} - \mathbf{f}_T) \quad (9)$$

Here, \mathbf{f}_T denotes the feature vector of target database image T in general for different image representation (e.g., global and semi-global). 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., it is not dependent on the scale of measurements. If the covariance matrix is the identity matrix then it is the same as the Euclidean distance [9].

Basically, at each iteration of relevance feedback, we generate several mean vectors and covariance matrices for each of the representation separately and use it in the distance measures. Finally, we obtain a ranked based retrieval by applying the fusion-based similarity function of equation (7). So, the above relevance feedback approach performs both the query point movement and similarity matching adjustment at the same time.

2.4 Integration of the textual and visual results

We have considered a pre-filtering and merging approach based on the text and image result lists obtained by the text retrieval after query expansion and image retrieval by applying adaptive distance measure (equation (9)) after the relevance feedback. In this multi-modal integration approach, combining the result of the text and image based retrieval is a matter of re-ranking or re-ordering of the images based on a weighted combination of scores from both the modalities. Instead of purely merging the results, we basically perform a pre-filtering step with the text query at first as it can generate results more closely to the user perceived semantics. The steps involved in the proposed interaction and integration approaches are as follows:

Step 1: Perform an initial text-based search with a query vector $\mathbf{D}_{q(0)}$ for a query topic $q(0)$ at iteration $i = 0$ and rank the associated images based on the ranking of the text (annotation) documents by applying S_{text} of equation(2).

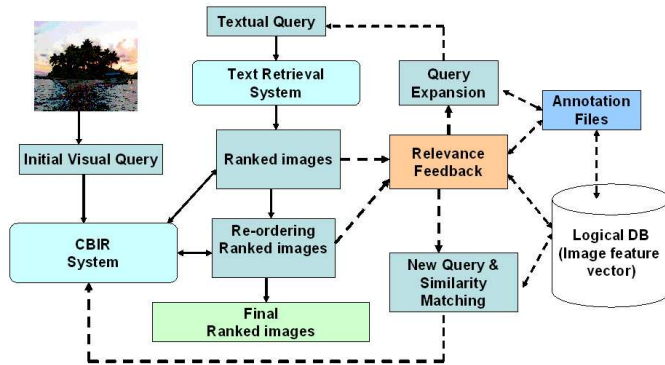


Figure 1: Process flow diagram of the integration approach

Table 1: Results of the ImageCLEFphoto Retrieval task

Run ID	Language	Mod	A/M	FB	QE	MAP
Cindi-Text-Eng	Eng	Text	Automatic	Without	No	0.1995
Cindi-TXT-EXP	Eng	Text	Manual	With	Yes	0.3749
Cindi-Exp-RF	Eng	Text+Image	Manual	With	Yes	0.3850

Step 2: Consider top $K = 30$ most similar images from the retrieval interface and obtain user feedback about positive or relevant images (e.g., associated annotation files) for the textual query expansion.

Step 3: Resubmit the modified query vector $\mathbf{D}_{q(i)}$ by re-weighting the keywords at iteration i . Continue the iterations by incrementing i , until the user is satisfied or the system converges.

Step 4: Perform visual only search on the result list of the first $L = 2000$ images obtained from step 3 with the initial query image $Q(0)$.

Step 5: Obtain the user feedback of the relevant images and perform the image only search with the new query $Q(i)$ at iteration i with equation (9) and equation (7). Continue the iterations by incrementing i , until the user is satisfied or the system converges.

Step 6: Aggregate the text and image based scores by fusing the similarity measures as:

$$S(Q, T) = w_{\text{text}}S_{\text{text}}(\cdot, \cdot) + w_{\text{image}}S_{\text{image}}(\cdot, \cdot) \quad (10)$$

where, $w_{\text{text}} = 0.7$ and $w_{\text{image}} = 0.3$ are selected for the experiment.

Step 7: Finally, rank the images in descending order of similarity values and return the top 1000 images.

Fig. 1 shows the process flow diagram of the proposed multi-modal interaction and integration approaches.

2.5 Analysis of the results

We have submitted three runs for the ad-hoc retrieval of the IAPR collection as shown in Table 1. In all these runs, English queries and example images are used as our initial source queries. In the first run with ID “Cindi-Text-Eng”, we performed only the automatic text-based search without any feedback as our base run. For the second and third runs with ID “Cindi-TXT-EXP” and “Cindi-Exp-RF” respectively, we performed manual feedback in the text only modality and in combination of the text and image modalities (with only one or two iterations for each modality) as discussed in the previous sections. From Table 1, it is clear that the MAP scores are almost

doubled in both the cases with the feedback and integration of text and image has achieved the best performance. In fact, these two runs ranked first and second in terms of the MAP score among the 157 submissions in the photographic retrieval task. Our group performed manual submissions using relevance judgement from the user, which along with the integration of the both modalities could be the reason for our good results.

3 Ad-hoc retrieval from medical image collections

For the ah-hoc image retrieval task in the medical collections (e.g., CaseImage, PEIR, MIR and PathoPic datasets), we have experimented with a similar cross-modal approach performed for photographic retrieval. However, for the text-based indexing and search, we have utilized the Lucene search engine [10], an open source project under the Apache software foundation. We have also performed a different query expansion and merging algorithm to obtain a text-based result list and finally merge with the image-based result list to obtain the final ranked result by applying similar weighting scheme as discussed in section 2.4 for photographic retrieval.

To use the textual information for image retrieval in the medical collections, each image has to be attached to at least one (possibly empty) text document. The text-based indexing is started by extracting keywords from the XML documents by parsing them using Xerces2 Java Parser, an open source project under the Apache software foundation. Every element of the XML document is indexed as a separate field in Lucene. Separate fields make it easier to search for the contents based on criteria or simply searching on all the indexed elements. Before indexing, the stop words (we have added additional domain specific stop words to the list) are removed from the description of the elements. Once the index creation process is completed, the keyword-based searching can be performed using the Lucene API.

For content-based indexing, we use the same approach as described in section 2 for the photographic collection. However, we have extracted a low-resolution scaled-specific image feature in addition to the global, semi-global and local region-specific features. Since, images in the different medical collections vary in sizes, resizing them into a thumbnail of a fixed size might reduce some noise due to the artifacts presents in the images, although it may introduce distortion. These approaches are extensively used in face or finger-print recognition and have proven to be effective. For the scaled-based feature vector $\mathbf{f}^{\text{Scaled}}$, each image is converted to a gray-level image and down scaled to 64×64 regardless of the original aspect ratio. Next, the down-scaled image is partitioned further with a 16×16 grid to form small blocks of (4×4) pixels. The average gray value of each block is measured and concatenated to form a 256-dimensional feature vector. By measuring the average gray value of each block, it can cope with global or local image deformations to some extent and adds robustness with respect to translations and intensity changes.

We also utilize the Euclidean distance measure to compare Q and T for $\mathbf{f}^{\text{Scaled}}$ and the fusion-based similarity function is slightly adjusted due to the added scaled-specific feature as follows:

$$S_{\text{image}}(Q, T) = w_g S_{\text{global}}(Q, T) + w_{\text{sc}} S_{\text{scaled}}(Q, T) + w_{\text{sg}} S_{\text{semi-global}}(Q, T) + w_l S_{\text{local}}(Q, T) \quad (11)$$

For the medical retrieval experiments, the weights are adjusted as $w_g = 0.4$, $w_{\text{sc}} = 0.2$, $w_{\text{sg}} = 0.25$ and $w_l = 0.15$.

3.1 Query expansion and integration of the results

A search process can start either by entering some text (query topic) in the text field or by providing a query image (e.g., “query-by-example”) to the system. If the user starts a keyword-based search process, then the system will search the index, find the XML documents where those keywords occur based on the similarity matching on the query and document vectors, and finally retrieve the images corresponding to the XML documents. Resulting images are then displayed as sorted in descending order of the similarity scores of the associated XML documents.

If the user starts the search process with the visual approach (i.e. “query by example”), various low-level image features will be computed on-line and the resulting images will be displayed sorted

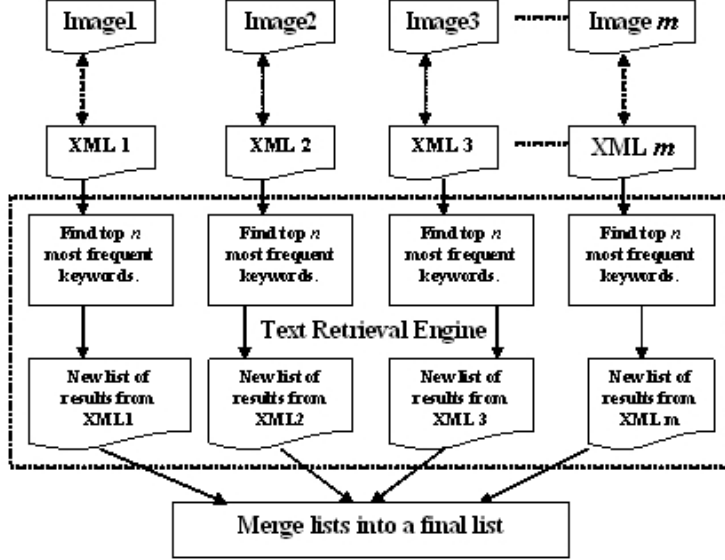


Figure 2: Query expansion and merging approach

by similarity score obtained from equation (11). After the initial search, user can make use of the relevance feedback system, which can work on both the text and image modalities simultaneously to display the results in the subsequent passes as discussed in section 2.3. After obtaining the initial results, user can select the relevant images as a positive feedback to the system which indicates the type of images the user is looking for. The system then runs two separate queries on the text and image-based systems for the selected feedback images.

For the query expansion in the text-based system, the system finds the corresponding XML documents from the positive feedback images. Next, it extracts the top n most frequent keywords from each XML documents. Hence, in the next iteration of RF, user will submit separate new queries to the system using the new keywords found in each document. This will result in m different lists of results where m is equal to the number of documents sent as positive feedback. After getting m separate lists of results, we merge these into a single list and display the text-based result.

Merging of the results is based on the assumption that if a particular image is occurring in most of the lists, then it should have a higher rank or priority than images less frequent in the lists. So, we upgrade the rank of this image by increasing the average similarity score based on how many lists contain that image. For example, if there are 10 lists of results, and a particular image presents in 8 of the lists, we will add more weight to this image than the image which occurs in 3 lists. More technically, if the image presents in all the list then a boost of 0.3 as additional score will be given to that image score. If the image exists in 50% or above of the lists then a boost of 0.2 and for less than 50%, a boost of 0.1 is given. For the images exist only once among all the lists, no boost is provided but the images are added to the list of the final result with their original score. After this, the list is sorted with the new scores and displayed as the final text-based result list as shown in the Fig. 2. The different boosting scores are selected by experimenting on a small sample database, which provided better results.

When the content-based system receives the list of positive images as relevance feedback, we can perform similar query point movement and similarity measure adjustment techniques as described in section 2.3, to return a new image-based result list on the text-based pre-filtered images. Once we have separate lists of results (e.g., one from the text-based system and another from the image-based system), we merge the lists using the similar weighting scheme as described in section 2.4 for the photographic image retrieval.

Table 2: Results of the Medical Retrieval task

Run ID	Topic	System	MAP	R-prec	B-pref
CINDI-Fusion-Visual	Automatic	Visual	0.0753	0.1311	0.166
CINDI-Visual-RF	Feedback	Visual	0.0957	0.1347	0.1796
CINDI-Text-Visual-RF	Feedback	Mixed	0.1513	0.1969	0.2397

3.2 Analysis of the results

We have submitted three runs for the ad-hoc medical retrieval as shown in Table 2. In all these runs, English queries and example images are used as our initial source queries. In the first run with ID “*CINDI-Fusion-Visual*”, we performed only the automatic visual only search without any feedback. Our group ranked first in this run category (automatic+visual) based on the MAP score (0.0753) out of five different groups and 11 submissions. For the second run with ID “*CINDI-Visual-RF*”, we performed the manual feedback in the image only modality. For this category (e.g., visual only run with RF), only our group has participated this year and achieved better MAP score (0.0957) than without RF as shown in Table 2. For the third run with ID “*CINDI-Text-Visual-RF*”, we performed the manual feedback in both the modalities and merge the result lists as discussed in the previous section. For this category (e.g., mixed with RF), we have achieved a moderate MAP score of 0.1513. From the scores, it is clear that combining both modalities is far better than using only a single modality (e.g., only image).

4 Automatic annotation tasks

The aim of the automatic annotation task is to compare the state-of-the-art approaches to image classification and annotation and to quantify their improvements for image retrieval. We investigate a supervised learning-based approach to associate the low-level image features with their high-level semantic categories for the image categorization or annotation of the medical (IRMA) and object (LTU) data sets. Specially, we explore the classifier combination approach of several probabilistic multi-class support vector machine (SVM) classifiers. Instead of using only one integrated feature vector, we utilize the features at the different levels of image representation as inputs to the SVM classifiers and use several classifier combination approaches to predict the final image category as well as probability or membership score of each category as image annotation.

4.1 Probabilistic multi-class SVM with pairwise coupling

SVM is an emerging machine learning technology that has already been used successfully for image retrieval and classification purposes [13]. It performs classification between two classes by finding a decision surface that is based on the most informative points of the training set. Briefly, one can say that SVM constructs a decision surface between samples of two classes, maximizing the margin between them. SVM was originally designed for binary classification problem. A number of methods have been proposed for extension to multi-class problem to separate L mutually exclusive classes essentially by solving many two-class problems and combining their predictions in various ways [14]. In the experiments, we utilize a multi-class classification method by combining all pairwise comparisons of binary SVM classifiers, known as *one-against-one* or pairwise coupling (PWC) [14]. PWC constructs binary SVM’s between all possible pairs of classes. Hence, this method uses $L*(L-1)/2$ binary classifiers, each of which provides a partial decision for classifying a data point. During the testing of a feature vector \mathbf{f} , each of the $L*(L-1)/2$ classifier votes for one class. The winning class is the one with the largest number of accumulated votes. Although the voting procedure requires just pairwise decisions, it only predicts a class label [16]. However, to

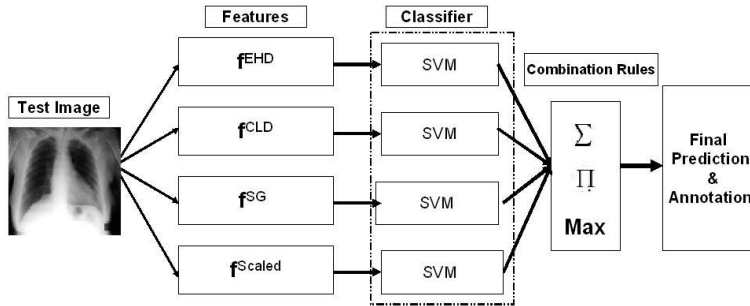


Figure 3: Block diagram of the classifier combination process.

annotate or represent each image with a category specific confidence score, probability estimation is required. In our experiments, the probability estimation approach in [14] for the multi-class classification by PWC is utilized. In this context, given the observation or feature vector \mathbf{f} , the goal is to estimate the posterior probability as

$$p_k = P(y = k | \mathbf{f}), k = 1, \dots, L \quad (12)$$

4.2 Multiple SVM classifier combination

The development of a multiple expert or classifier combination based system has received increasing attention and has been a popular research topic. In general, classifier combination is defined as the instances of the classifiers with different structures trained on distinct feature spaces [15]. Feature descriptors at different levels of image representation are in diversified forms and often complementary in nature. It is rather unwise to concatenate them together to form a single feature vector as the input for a single classifier. Hence, multiple classifiers are needed to deal with the different features, which results in a general problem of how to combine those classifiers with different features to yield the improved performance.

In the experiments, we consider expert combination strategies of the SVM classifiers with different low-level features as inputs based on three popular classifier combination rules (e.g., sum, product and max rules) [15]. Since the outputs of the classifiers are to be used in combination, the confidence or membership scores from the probabilistic SVM's in the range of $[0, 1]$ for each category serve this purpose. In these combination rules, a priori probabilities are assumed to be equal and the decision is made by the following formula in terms of the a posteriori probabilities yielded by the respective classifiers:

$$P_k^{\text{combine}}(y = k | \mathbf{f}) = \frac{P_k^{\text{combine}}}{\sum_{k=1}^L P_k^{\text{combine}}}, k = 1, \dots, L \quad (13)$$

Here, P_k^{combine} is the combined output of the classifiers about the likelihood of the sample vector \mathbf{f} belonging to the category $k \in L$. P_k^{combine} and is obtained by using the following two combination rules [15]:

In product rule, it is assumed that the representations used are conditionally statistically independent, where R experts or classifiers are combined as follows

$$P_k^{\text{combine}} = \prod_{m=1}^R P(y = k | \mathbf{f}^m) \quad (14)$$

where, $P(y = k | \mathbf{f}^m)$ denotes the posterior probability of the class k on the input \mathbf{f}^m for the classifier m . Similarly, for the sum and max rules, it can be stated as follows:

$$P_k^{\text{combine}} = \sum_{m=1}^R P(y = k | \mathbf{f}^m) \quad (15)$$

Table 3: Performance of the object annotation task (LTU dataset)

Run ID	Feature	Method	Error rate(%)
Cindi-SVM-Product	CLD+EHD+Semi-global	SVM (Product)	83.2
Cindi-SVM-SUM	CLD+EHD+Semi-global	SVM (Sum)	85.2
Cindi-SVM-EHD	EHD	SVM	85.0
Cindi-Fusion-Knn	CLD+EHD+Semi-global+Local	K-NN	87.1

Table 4: Performance of the medical annotation task (IRMA dataset)

Run ID	Feature	Method	Error rate(%)
cindi-svm-product	CLD+EHD+Scaled+Semi-global	SVM (Product)	24.8
cindi-svm-sum	CLD+EHD+Scaled+Semi-global	SVM (Sum)	24.1
cindi-svm-max	CLD+EHD+Scaled+Semi-global	SVM (Max)	26.1
cindi-svm-ehd	EHD	SVM	25.5
cindi-fusion-KNN9	CLD+EHD+Scaled+Semi-global+Local	K-NN	25.6

$$P_k^{\text{combine}} = \max_{m=1}^R P(y = k | \mathbf{f}^m) \quad (16)$$

The sum rule is developed under stricter assumptions than the product rule. In addition to the conditional independence assumption in the product rule, the sum rule assumes that the probability distribution will not deviate significantly from the a priori probabilities [15]. The multi-class SVM classifiers as experts on different feature descriptors as described in section 2.2 for retrieval, are combined with the above rules and finally classify the image to the category with the highest obtained probability value and annotate the images with the probability or membership scores as shown in the process diagram in Fig. 3.

4.3 Analysis and results of the runs

To perform SVM-based classification, we utilize the LIBSVM software package [16]. For the training of both the data sets, RBF kernel functions are utilized with different kernel γ and cost C parameters found out experimentally with 5-fold cross validation (CV).

We have submitted four runs for the object annotation task as shown in Table 3. First three of these runs are experimented with the proposed multi-class SVM and classifier combination approach with different feature inputs and the last run with ID “*Cindi-Fusion-Knn*” is experimented with a K-NN (K=9) classifier by using the fusion-based similarity matching function in equation (7). Our best run (e.g., “*Cindi-SVM-Product*”) in this task ranked third among all the submissions. Although the accuracy rate is much lower at this moment due to the complexity of the dataset in general.

For the medical annotation task, we have submitted five runs as shown in Table 4. First four of these runs are experimented with the proposed multi-class SVM and classifier combination approach and the last run with ID “*cindi-fusion-KNN9*” is experimented with a K-NN (K=9) classifier by using the fusion-based similarity matching function in equation (11). Our best run (e.g., “*cindi-svm-sum*”) in this task ranked 13th among all the submissions and 6th among all the groups.

5 Conclusion

This report has examined the image retrieval and annotation approaches of CINDI research group for ImageCLEF 2006. We have participated in all the four sub-tasks and submitted several runs with different combination of methods, features and parameters. We have experimented with a cross-modal interaction and integration approach for the retrieval of the photographic and medical image collections and a supervised classifier combination-based approach for the automatic annotation of the object and medical datasets. The analysis and the results of the runs are discussed in this paper.

References

- [1] P. Clough, M. Grubinger, T. Deselaers, A. Hanbury, H. Müller, “Overview of the ImageCLEF 2006 photo retrieval and object annotation tasks,” *CLEF working notes*, Alicante, Spain, Sep., 2006.
- [2] H. Müller, T. Deselaers, T. Lehmann, P. Clough, W. Hersch, “Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks”, *CLEF working notes*, Alicante, Spain, Sep., 2006.
- [3] R. Baeza-Yates and B. Ribiero-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.
- [4] R. J. Mooney, “Intelligent Information Retrieval and Web Search ”, Online Courseware, University of Texas, Austin, USA, available at <http://www.cs.utexas.edu/users/mooney/ir-course/>
- [5] 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.
- [6] M. M. Rahman, B.C. Desai, P. Bhattacharya, “A Feature Level Fusion in Similarity Matching to Content-Based Image Retrieval”, *Proc. 9th Internat Conf. Information Fusion*, 2006.
- [7] B. S. Manjunath, P. Salembier, T. Sikora, (eds.) *Introduction to MPEG-7- Multimedia Content Description Interface*, John Wiley Sons Ltd. pp. 187-212, 2002.
- [8] S. Aksoy, R. M. Haralick, “Texture Analysis in Machine Vision”, *Chapter Using Texture in Image Similarity and Retrieval, Series on Machine Perception and Artificial Intelligence.*, World Scientific, 2000.
- [9] K. Fukunaga, *Introduction to Statistical Pattern Recognition* , 2nd ed. Academic Press, 1990.
- [10] Lucene search engine, available at <http://lucene.apache.org/java/docs/>
- [11] Y. Rui, T. S. Huang, “Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval”, *IEEE Circuits Syst. Video Technol.*, vol. 8, 1999.
- [12] J. Friedman, “Regularized Discriminant Analysis”, *Journal of American Statistical Association*, vol. 84, pp. 165–175, 2002.
- [13] V. Vapnik *Statistical Learning Theory*, New York, NY, Wiley; 1998.
- [14] T. F. Wu, C. J. Lin, R.C. Weng, “Probability Estimates for Multi-class Classification by Pairwise Coupling”, *J Machine Learning Research* vol. 5, pp. 975–1005, 2004.
- [15] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, “On combining classifiers”, *IEEE Trans Pattern Anal Machine Intell* vol. 20(3), pp. 226–239, 1998.
- [16] C. C. Chang, C. J. Lin, “LIBSVM : a library for support vector machines”, Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, 2001.