IBM Research at ImageCLEF 2013 Medical Tasks

Mani Abedini[1], Liangliang Cao[3], Noel Codella[3], Jonathan H. Connell[3], Rahil Garnavi[1], Amir Geva[2], Michele Merler[3], Quoc-Bao Nguyen[3], Sharathchandra U. Pankanti[3], John R. Smith[3], Xingzhi Sun[1], and Asaf Tzadok[2]

[1] Level 5, 204 Lygon St. Carlton
Victoria 3053, Australia
{mabedini,rahilgar,xingsun}@au1.ibm.com
http://www.research.ibm.com/labs/australia

[2] Mount Carmel
Haifa, 31905, Israel
{asaf,geva}@il.ibm.com
http://www.research.ibm.com/labs/haifa

[3] 1101 Kitchawan Rd.
Yorktown Heights NY 10598, USA
{liangliang.cao,jconnell,nccodell,mimerler,quocbao,sharat,jsmith}@us.ibm.com
http://www.watson.ibm.com

Abstract. In this paper we present the modeling strategies that were applied by the IBM Research team to the medical modality classification, retrieval and compound figure separation tasks of ImageCLEF 2013.
We present our methods for each task and discuss our submitted textual, visual, and mixed runs, as well as their results, the use of external resources and human supervision.
The key components of our modality classification submissions were: 1)fusion of multiple low level image descriptors extracted at different spatial granularities 2) use of pre-existing medical categories classifiers trained from web sources 3) pseudo-probabilistic analysis of modality-specific derived text patterns and 4) multiple fusion strategies to combine visual and textual information.
For the case based retrieval task, we applied topic modeling on top of text extracted from the full set of Pubmed articles, using three different corpuses to guide an expansion: one from UMLS keywords, one from WordNet keywords, and one from the union of the previous two. Retrieval was performed using Lucene indexing on top of such representations. For the image based retrieval task, we adopted visual image similarity based on CHI square distance between low level visual descriptors.
In the compound figure separation task we tried a combination of two approaches, one based on a connected components analysis in a binarized image, the other adopting the common notation of subfigures using text.

Keywords: Medical Modality Classification and Retrieval, Muliclass SVM, Text Patterns, Topic Modeling, Figure Spearation

1 Introduction

ImageCLEF medical track is the cross-language image retrieval track of the Cross Language Evaluation Forum focused on the analysis of medical images. The ImageCLEF medical track 2013 consists of four tasks: modality classification, case-based retrieval, image-based retrieval and compound figure separation, which is a new task introduced this year. We participated to all tasks, although we submitted official runs only for the first three.

The remaining of the paper id organized as follows: Section 2 describes the visual and text based approaches utilized for the modality classification task, in Sections 3 and 4 we cover the details of our case based and image based retrieval approaches, respectively. We present our method for compound figure separation in Section 5 and finally we draw some conclusions and possible future directions in Section 6.

2 Modality Classification

The key components of our modality classification submissions were: 1)fusion of multiple low level image descriptors extracted at different spatial granularities 2) use pre-existing medical categories classifiers trained from web sources 3) pseudo-probabilistic analysis of modality-specific derived text patterns and 4) multiple fusion strategies to combine visual and textual information. In the following we describe each component in detail.

2.1 Feature Extraction

Visual Descriptors (Runs 3,4,5,6,7,8,9,10)

All our experiments were based on a set of low-level visual features extracted at different spatial granularities. We selected a subset of the most useful features with the use of a 80%-20% train-validation split of the whole training set. The spatial granularities adopted were as follows:

- **Global**: Feature extracted from entire image.
- **Grid(7)**: 5x5 (7x7) image grid, with feature vector extracted from each grid block and concatenated. Increases dimensionality by factor of 25 (49).
- **Layout**: 5 image regions including the center and the 4 quarters.
- **Pyramid(3)**: spatial pyramid with global as first level and 2x2 (3x3) image grid as second level.

The pool of visual features we employed is a combination of global and local descriptors:

- **Color Histogram**: global color distribution represented as a 166-dimensional histogram in HSV color space.
- **Color Correlogram**: global color and structure represented as a 166-dimensional single-banded auto-correlogram in HSV space using 8 radii depths.

- **Color Moments**: localized color extracted from a 5x5 grid and represented by the first 3 moments for each grid region in Lab color space as a normalized 225-dimensional vector.
- **Wavelet Texture**: localized texture extracted from a 3x3 grid and represented by the normalized 108-dimensional vector of the normalized variances in 12 Haar wavelet sub-bands for each grid region.
- **Edge Histogram**: global edge histograms with 8 edge direction bins and 8 edge magnitude bins, based on a Sobel filter (64-dimensional).
- **GIST**: describes the dominant spatial structure of a scene in a low dimensional representation, estimated using spectral and coarsely localized information. We extract a 512 dimensional representation by dividing the image into a 4x4 grid, we also extract histograms of the outputs of steerable filter banks on 8 orientations and 4 scales.
- **Local Binary Patterns (LBP)** [1]: extracted from the greyscale image as a histogram of 8-bits local binary patterns, each of which is generated by comparing the greyscale value of a pixel with those of its 8 neighbors in circular order, and setting the corresponding bit to 0 or 1 accordingly. A pattern is called uniform if it contains at most two bitwise transitions from 0 to 1. The final histogram for each region in our granularity contains 59 bins, 58 for uniform patterns and 1 for all the non-uniform ones.
- **Image Type**: a set of global image statistics: mean saturation, hue entropy, variance and switches, quantized color entropy and switches
- **Image Stats**: a series of statistics on the image, namely: aspect ratio, isWhite, isBlack, isAllBW, isColor, entropy, variance, minum value, maximum value, mean, median, standard deviation, central moments, average energy of of first level 2D wavelet decomposition subbands, skin color, number of unique colors in quantized color space.
- **Curvelets**: computes the magnitude of a spatial pyramid in Fourier space, under the polar coordinate system, across all 3 color channels red, green, blue, in addition to a grayscale color channel. Pyramid levels in the radial dimension consist of 1, 2, 4, and 8 partitions. For each of these partitions, we construct a pyramid in the angular dimension, of partitions 1, 2, 4, 8, 16, and 32 segments. Due to the property of image symmetry in Fourier space, only the top half of the polar Fourier circle is sampled for the feature vector.
- **Maxi Thumbnail Vector**: the concatenated RGB pixel values of the image after down-sampling to 24x24 dimensions.
- **SIFT[2] AM**: SIFT descriptor extracted around Harris Laplace interest points. Each keypoint is described with a 128-dimensional vector containing oriented gradients. We obtain a visual words dictionary of size 1000 by running K-means clustering on a random sample of approximately 300K interest point features, we then represent each image with a histogram of visual words. We extracted two codebooks, starting from two different random samples of points. We used soft assignment following Van Gemert et al. [3] using $\sigma = 90$. This descriptor was extracted using the executable publicly available from the University of Amsterdam [4]. We extracted also varia-

tions of the SIFT descriptor in different color spaces, namely rgb, hsv and opponent channels.
– **SIFT VLFEAT**: similar to SIFT AM. with the difference that the local descriptor were extracted using the VLFEAT library [5].

Medical Semantic Model Vector (Runs 5,6,9,10). In addition, 121 medical semantic concept classifiers were constructed from training data collected from various sources (IRMA, TCIA, JSRT, Web Crawl) using the IMARS framework 2.2. These classifiers cover a range of radiological modalities, body regions, views, and some instances of disease pathology. They do not cover non-radiological image categories. Each of these concept classifiers was scored against the ImageCLEF training and test data, and concatenated into a 121 dimensional feature vector used in some experiments, following a similar framework adopted by our group for video event retrieval [6]. From the official test set performance reported in Figure 1 we can notice how the Medical Semantic Model Vector is the single best performing visual descriptor.

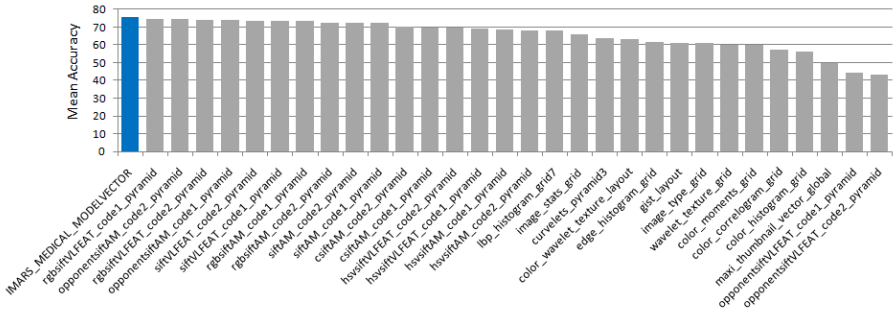


Fig. 1. Test performance of multiclass χ^2 SVM based on individual visual descriptors. The best performing visual feaure is the IMARS Medical Semantic Model Vector, with an average accuracy of 75.8%.

Textual Descriptors We generated three different types of textual descriptors, described in detail in the following.

– **Modality Taylored Keywords** (Runs 1,7,8,9): The text-based classifier built on top of this representation generates a likelihood score for each modality based on the presence or absence of a number of key words. There are over 400 such text patterns which can be either full words, fragments of words, or multi-word phrases. The vocabulary terms were hand selected by perusing roughly half of captions in the training set. Between 2 and 51 patterns were selected for each modality then combined into one big feature list. Related phrases such as *fluorescent*, *immunofluorescence*, and *Alexafluor* were

merged to variablized patterns such as **fluor**. The asterisks at the front and/or back match an arbitrary number of characters up to the first token delimiter (e.g. space). This makes the statistics more robust since there are more matching examples found for the variabilized pattern than for the original, more specific terms. Also important is the fact that patterns with all capital letters were only matched to text that was fully capitalized. Otherwise a pattern like **PET* could potentially match many irrelevant words. The number of hits (or an absence of a hit) for each term is then weighted by a pseudo-probablistic model derived from the known modalities of the training examples. The conditional probability of seeing a term given a particular modality is divided by that term's background probability. This is used as a boosting factor for the modality's likelihood and is applied for each occurrence of the term found in the probe caption. To keep the small sample size from provoking overfitting of probabilities, an estimate is made of the range of expected probabilities for each expression. Only if the low estimate for the conditional probability exceeds the high estimate for the background (or vice versa) is the term assigned a boosting factor. A similar set of factors is generated for the complete absence of a term. For instance, angiograms usually mention *artery*. If this term is missing then the probe is less likely to be an angiogram. Finally the factors for the presence and absence of all 400+ terms in the probe caption are combined to give a score for each modality, and the highest scoring class is picked as the winner.

– **Ontology Based Vocabulary** (Run 2): Given words used in the medical articles can not be found in one single ontology, we applied the textual retrieval task based on several ontologies from a general lexical ontology (WordNet [7]) to medical specific domains medical knowledge-bases such as the Unified Medical Language System (UMLS) Metathesaurus [8], SNOMED-CT and RxNorm dictionaries. We built a NLP pipeline that consist of WordNet lexical relations [9], the Clinical Text Analysis and Knowledge Extraction System (cTAKES) and the Yale cTAKES Extensions (YTEX [10]). We applied the word-sense disambiguation and sliding window based part-of-speech to identify the relationships among words in the context of medical articles and the types of clinical named entities such as drugs, diseases, and symptoms. In order to the support the cross-validation and the fusion, the medical articles were indexed using Lucene technology into several categories such as titles, abstracts, captions and different domains such as general lexical ontology (WordNet) and medical knowledge-bases for later retrieval. This approach allows performing effective searches on a multidimensional vector-space with a FrequencyInverse Document Frequency (TF-IDF) weighting. This statistical measure allowed us to conduct several search experiments to evaluate the importance of the generic terms not related to medical domain and/or clinical named entities to a document in a corpus. The importance increases with the frequency of the word in the specific document but decreases with the frequency of the word in the whole corpus.

– **Bag of words** (Run 10): This method is one of the common approaches to extract features from a text (image captions in our case), which incorporates counting the frequency of the words appeared in the captions. Then each caption is represented by a term vector. The term-space vector is associated with a dictionary. Two dictionaries have been used in our approach: Manually generated and Automatically generated dictionaries (MGD and AGD respectively). The MGD is manually generated from a list of keywords nominated by a subject matter expert, which contains 123 words. The AGD has been generated by extracting nouns from the training image captions using python NLT toolkit. In order to reduce the size of dictionary we applied two main modification. First, all words are grouped based on the morphology analysis, where each groups is represented by a main morpheme in the dictionary. For example, any plural form is transformed into singular form, and all synonyms are replaced with a single noun. Second, the words should be appeared in the captions at least twice. AGD contains 8930 words.

2.2 Classfiers Modeling and Fusion

Since the set of descriptors, either visual or textual, we extracted was quite large, this year we experimented with a variety of multiclass modeling and fusion strategies, summarized in the following.

IMARS modeling (Runs 5,9)

The IBM Multimedia Analytics and Retrieval System (IMARS) is designed as a system to train 1-vs-all classifiers using late fusion of features. The system is broken down into two primary components: unit model training, and ensemble model fusion. Data is broken into two partitions for each stage: Learning (typically 80%) and Validation (typically 20%). Resultant 1-vs-all classifiers are combined into a multiclass system by choosing the label that produces the maximum score.

In the unit model training stage, a single feature type is selected over a random subset of data from the Learning data partition. A model is then trained for this "bag", or cross-product, of feature and data. For the purposes of ImageCLEF, we used SVMs. Kernel parameter selection was done using grid-search and 2-fold cross-validation within the bag, optimizing accuracy. In order to compensate for the possibility of data imbalance, either bags are forced to take balanced samples of data (positive and negative), or a proprietary variant of the SMOTE algorithm is employed, that has been shown to improve performance over the original SMOTE algorithm under certain conditions of training SVMs. Data balance in this experiment is important, since imbalance can shift resultant SVM scores, yielding error when using a maximum-value operator to convert 1-vs-all classifier scores to multiclass.

Once unit models have been trained over all features and data, they are combined using a forward model selection process that optimizes the accuracy of the final 1-vs-all ensemble classifier.

Two level SVM + Kernel Approximation (Runs 4,6,8)

In ImageCLEF2012, we tried the linear kernel approximation method to fuse multiple features [12] which obtained 61.5% accuracy in the medical modality test. The idea of our kernel approximation based is very simple: first map each feature into a higher dimension space using explicit kernel mapping [13], and train a linear model with all the concatenated features using LibLinear [14]. Figure 2 illustrates the fusion method in ImageCLEF 2012.

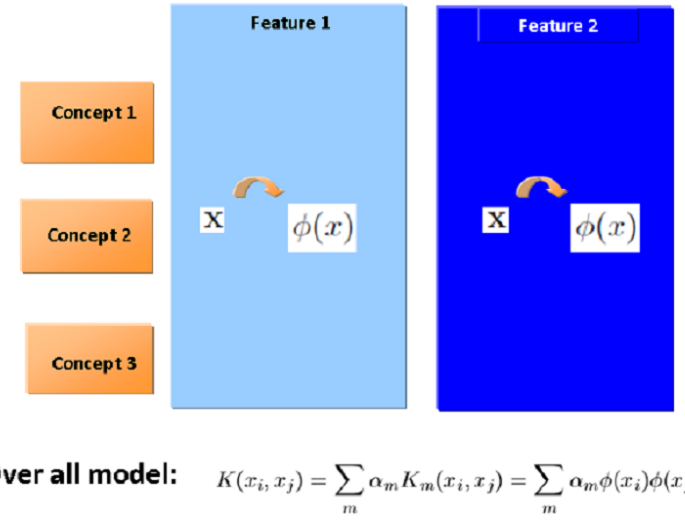


Fig. 2. Linear kernel approximation based fusion method in ImageCLEF 2012.

Although the kernel approximation method showed good performance, however, it has difficulties if the feature vectors are of very high dimensions or if number of features grows very big. In such scenario, since each image will be represented by a ultra-high dimensional vector, we will meet problems in loading all the data into the memory and/or normalize all the features appropriately.

In ImageCLEF 2013, we develop a new method which can be scalable in fusing many features and also high dimensional features. Our work is motivated by our 2012 work, our previous work on kernel fusion [15]. the deep learning framework [16]. Figure 3 illustrates our new method. Intuitively, our new method can be divided into four steps: (1) Represent a image with multiple features, (2) Map each feature in to high dimensional space where the linear product can approximate additive kernels, (3) Design a two level model which can handle a large amount of features and combine them into an efficient fusion model, and (4) Develop an incremental way for model selection on the training set. Our single run gets an accuracy of 80.87% using 24 features, 80.13% using 11 features, which is used as the component of with the best performed submission in ImageCLEF 2013 medical modality classification task.

Meta Classifiers (Run 10)

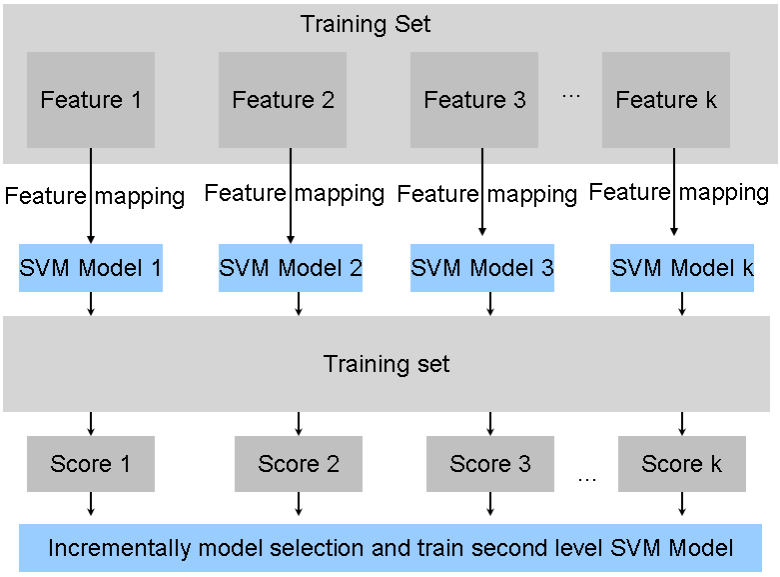


Fig. 3. New kernel approximation with two level SVM fusion.

Meta-learning is a strategy to learn from learned knowledge [17]. In conjunction with the ensemble learning methodology, an ensemble meta-classifier is an ensemble classifier on top of a collection of classifiers. Inspired from this methodology, we implemented another level of supervised learning classification for combining the results of fusion models. In IMARS, a straight forward majority vote by considering the maximum score, has been implemented which enforce competition among fusion models. In this approach we built a collaboration model to combine the fusion models predictions.

The input of the ensemble meta-classifier is a vector of SVM-fusion model scores from visual and text classifiers. Several machine learning classification methods have been explored in our experiments including: Decision Tree, Support Vector Machine (RBF Kernel, Poly kernel, Normalized Ploy kernel and Puk kernel), Random Forest, Logistic Model Tree (LMT), Naive Bayesian. The Weka [18] machine learning tool has been used for meta-classification purpose.

Early (Kernel) and Late Fusion (Runs 3,4,6,7,8,9)

We also experimented with standard feature/method fusion methods:

- **Kernel Fusion**: consists of a point-wise average pooling over the kernel matrices produced by each descriptor. The Multiclass SVM is then learned on top of the aggregate matrix adopting the 1 vs 1 plus majority voting scheme. Given this implementation, kernel fusion is equivalent to early fusion.
- **Late Fusion**: consists of a max pooling operator over the predictions of the models learned from different strategies for each test image.

In particular late fusion was used to produce most of our mixed runs.

2.3 Submitted Runs

In the following we present a summary of the ten modality classification runs that were submitted for evaluation.

- **Run 1**: textual classifier based on pseudo-probabilistic analysis on top of the Modality Taylored Keywords.
- **Run 2**: textual classifier based on Topics extracted from the ontology based vocabulary Pubmed articles set, with UMLS and WordNet guided expansion. Modality based indexing based on the derived topics was performed using the Lucene indexing tools.
- **Run 3**: visual multiclass χ^2 SVM with kernel based fusion and 1 vs.1 plus majority voting scheme, on top of all the visual descriptors with the exception of the Medical Semantic Model Vector.
- **Run 4**: late fusion of three visual multiclass classification methods based on a validation based selection from the pool of visual Descriptors, with the exception of the Medical Semantic Model Vector.
- **Run 5**: Random non-shared subspace bagging + Forward model selection, with the addition of a Semantic Model Vector, trained from an IBM medical image taxonomy, as one of the features
- **Run 6**: Same as run4, with the addition of a Semantic Model Vector, trained from an IBM medical image taxonomy, as one of the features
- **Run 7**: Late fusion of visual from run 3 + text from run 1
- **Run 8**: Late fusion of visual from run 4 + text from run 1
- **Run 9**: Super fusion within random non-shared subspace bagging + forward model selction of all visual and textual features
- **Run 10**: This run is based on ensemble meta-classifier (see Section 2.2). First visual and textual analysis models have been trained. The visual classifiers are SVM fusion models based on IMARS (see Section 2.2) and the textual classifiers are SVM classifiers using on bag of words as input features. We allocated 80% of randomly selected training samples to train these classifiers. The other 20% of the image CLEF training samples has been used for meta-learning training phase. A wide range of different meta classifiers with various parameter values have been explored to tune the meta-classifiers running 10 fold cross validation. As a conclusion we found that LMT demonstrates more accurate performance in respect to others.

2.4 Results

We achieved top performance for each category of submissions: Textual, Visual and Mixed. Figure 4 presents our submitted runs in comparison to the submissions from other groups.

For the text based runs, we found Modality Taylored Keywords to work best 64.17% mean accuracy. We performed additional experiments with this modeling strategy after the submission deadline and found that when mentions in the

Run Name	Retrieval Type	Run Type	Additional Resources
IBM_modality_run1	Textual	Human Assistance	none
IBM_modality_run2	Textual	Automatic	Lucene, WordNet, USMLS Pubmed articles set
IBM_modality_run3	Visual	Automatic	none
IBM_modality_run4	Visual	Automatic	none
IBM_modality_run5	Visual	Automatic	IBM medical image taxonomy Semantic Model Vector
IBM_modality_run6	Visual	Automatic	IBM medical image taxonomy Semantic Model Vector
IBM_modality_run7	Mixed	Human Assistance	none
IBM_modality_run8	Mixed	Human Assistance	none
IBM_modality_run9	Mixed	Human Assistance	Lucene, WordNet, USMLS, Pubmed articles set, IBM medical image taxonomy Semantic Model Vector
IBM_modality_run10	Mixed	Human Assistance	IBM medical image taxonomy Semantic Model Vector

Table 1. Breakdown of Modality Classification submitted runs in terms of run type and use of additional training data besides the official training set.

article text are combined with the figure captions, the overall system accuracy rises to 69.6%.

For the visual based runs, we noticed that the best individual visual descriptor was the Medical Semantic Model Vector (as reported in Figure 1. As in the past year, the combination of multiple descriptors proved to be beneficial. This year we further combined different modeling strategies (IMARS modeling, two-level SVM, multiclass SVM Kernel fusion) starting from a common pool of descriptors, which provided a further improvement in the overall classification performance (80.79% in Run 4).

Finally, consistently with the results of other teams from past years, we found that textual and visual analysis complement each other and registered a slight performance improvement in the mixed runs which combine such multimodal information. After analyzing the complementary nature of the results in the confusion matrices of the different modalities, as shown in Figure 5, we believe that such multimodal fusion could lead to a more significant improvement than the one we obtained in the submitted mixed runs, where simple model selection or late fusion strategies were employed.

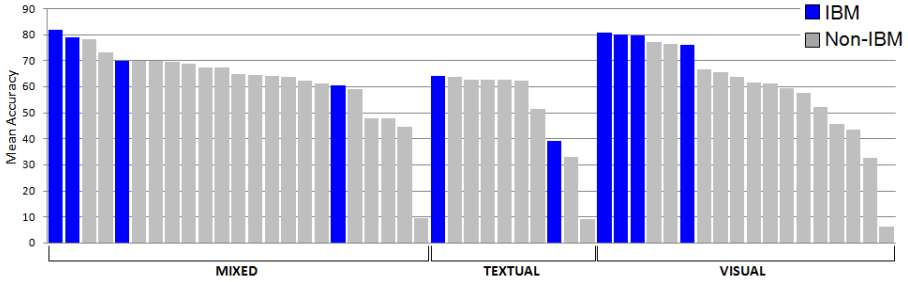
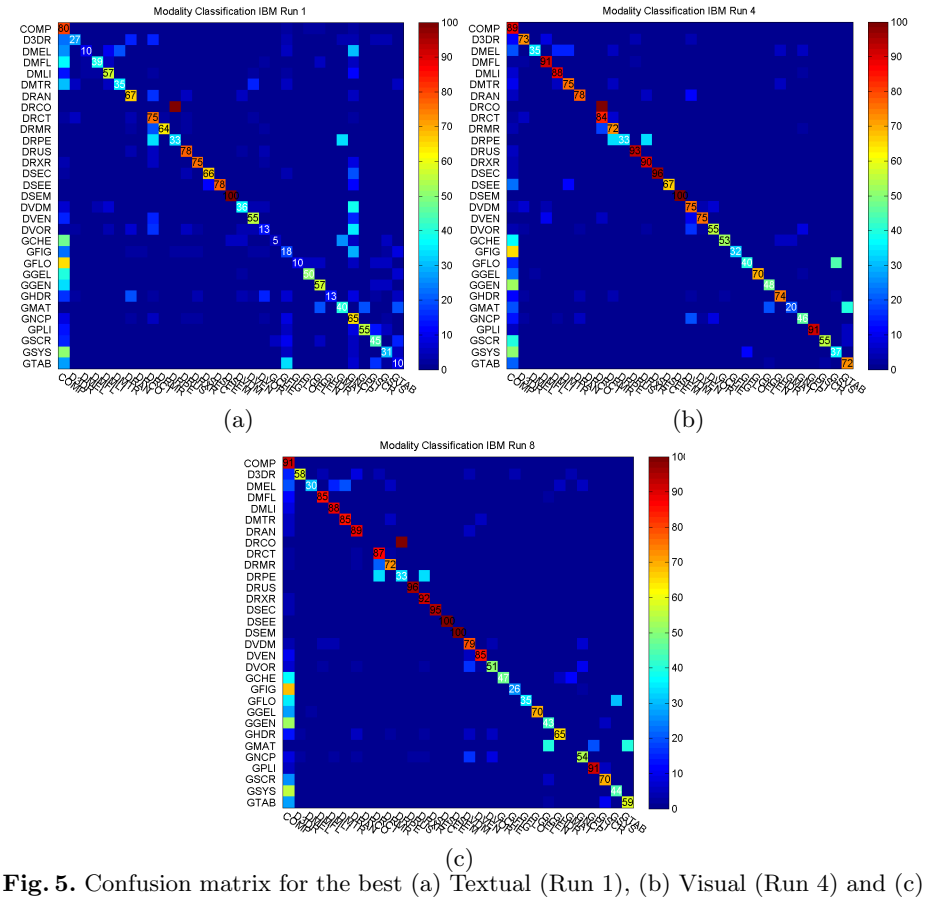


Fig. 4. Official Test results for the submitted Modality Classification runs. IBM runs are highlighted in blue and achieved top performances for each submission type (visual, textual, mixed), as well as overall best performance.



(a) (b)

(c)

Fig. 5. Confusion matrix for the best (a) Textual (Run 1), (b) Visual (Run 4) and (c) Mixed (Run 8) runs. Note the complementarity between errors the Textual and Visual runs. A more sophisticated fusion strategy could lead to further improvements in the Mixed runs.

3 Case Based Retrieval

3.1 Topic Modeling

In addition to the NLP pipeline and indexation techniques described in section 2.1, we found it is important to analyze the statistical structure of the corpus of medical documents in order to capture meaningful semantic patterns that can improve the process of classification and retrieval. Topic modeling [19] provides a method for learning the topics from a large text corpus and a topic can be defined as a collection of words that occur together frequently. We applied the topic modeling approach to identify meaningful patterns from the medical documents In our study, we used MALLET [20], an open-source toolkit to apply a Latent Dirichlet Allocation (LDA [21]), to detect the probability distribution $P(w|z)$ over words given topic z. Each medical document can be defined as a mixture of latent topics characterized by a multinomial distribution over words. In our experiments, we varied the number of topics ranging from 100 to 10,000 topics and used the Gibbs sampling and Bayesian estimation to assign the multinomial distributions over a set of words to each latent topic. Our goal aimed to reduce the terms that occur in a lot of topics that can lead to poor retrieval accuracy with non-relevant documents. They showed that the extracted topics captured meaningful structure in the document, consistent with the abstract and the title of the article.

We also explored the Jenson-Shannon divergence method to compute the multinomial distributions among the mixture topics that help to determine relevant articles for a given query.

In the process of indexing the topics, we separated the topics that are defined for titles, abstracts and captions and grouped the medical documents that share the same topics. One advantage of this approach is to be able to identify relevant documents to a particular query.

We incorporated the statistical distributions of the words in the topics into the TF-IDF weighting during the indexation of the topics. Consequently, we defined a weight for each $[term, topic]$ pair denoting the relative importance of the term to the topic and a weight for each $[topic, document]$ pair indicating the relative relationship of the topic in the document.

3.2 Submitted Runs

Our experiments were based on 75,000 articles including 300,000 images provided by the ImageCLEF dataset. Our NLP process analyzed each document by sliding the window based part-of-speech after tokenization, lemmatization and stopwords removal. The indexing process divided each article into several categories such as title, abstracts, full-text and captions. In addition, we also performed the indexing of 10,000 topics with each article document a mixture of a set of topics. We selected the number of topics that were most represented from the Ground-Truth and learned later that the size of topics and the distribution of

the mixture of topics for medical documents can impact the retrieval accuracy and performance.

Runs 1,2 and 3 employed 100,000 topics extracted from the UMLS keywords, WordNet keywords, and Mixed (UMLS + WordNet) vocaularies, respectively. Runs 4 to 6 follow the same framework, but we performs an automatic query expansion on top of the queries from Runs 1 to 3.

3.3 Results

We performed several runs by mixing the words from different ontologies as well as combinations of different categories such as titles, abstracts, full-text and captions. The results obtained by our submitted runs demonstrated the use of topic modeling and the combination of mixed ontologies can be effective for the top query hits such as P@10. We also learned by extending the number of topics during the training process with the Ground-truth can give us a better overall result but can affect the P@10 results.

Runid	**Retrieval type**	**MAP**	**M-MAP**	**bpref**	**P10**	**P30**
SNUMedinfo9	Textual	0.2429	0.1163	0.2417	0.2657	0.1981
IBM_case_run1	Textual	0.1573	0.0296	0.1596	0.1571	0.1057
IBM_case_run3	Textual	0.1573	0.0371	0.139	0.1943	0.1276
IBM_case_run6	Textual	0.1482	0.0254	0.1469	**0.2**	0.141
IBM_case_run2	Textual	0.1476	0.0308	0.1363	**0.2086**	0.1295
IBM_case_run4	Textual	0.1403	0.0216	0.138	0.1829	0.1238
IBM_case_run5	Textual	0.1306	0.0153	0.134	0.2	0.1276

Table 2. Case based retrieval results.

4 Image Based Retrieval

For the image based retrieval task, we adopted a purely visual based approach. The main goal was to estimate the basic retrieval power of individual descriptors which offered, by themselves, reasonable performance in the modality classification task.

The descriptors adopted were the same visual descriptors adopted for modality classification and described in detail in Section 2. Image similarity was based on the χ^2 square distance between such low level visual descriptors. When more than one image was part of a case, we picked either the max (runs named *nozero*) or average (*avg*) distance between the sets. The returned ranked list followed the decreasing distance scores between cases, without using any ad-hoc indexing schemes.

4.1 Submitted Runs

We submitted three runs, each based on a different descriptor:

- **IBM_image_run_Mnozero17**: based on the color HSVsiftAM descriptor extracted at pyramid granularity.
- **IBM_image_run_Mavg7**: based on the edge histogram descriptor extracted at grid granularity.
- **IBM_image_run_Mnozero11**: based on the regular siftAM descriptor extracted at pyramid granularity.

4.2 Results

The results of our runs are reported in Table 3, in comparison with the best performing run (DEMIR4). The results show that large room for improvement is left from our baseline runs, as the descriptors by themselves do not seem to be sufficiently powerful to provide acceptable performance. In particular a combination of multiple descriptors based ranked lists and the employment of more sophisticated indexing schemes could provide a significant improvement.

Runid	Retrieval type	MAP	M-MAP	bpref	P10	P30
DEMIR4	Visual	0.0185	0.0005	0.0361	0.0629	0.0581
IBM_image_run_Mnozero17	Visual	0.003	0.0001	0.0089	0.02	0.0105
IBM_image_run_Mavg7	Visual	0.0015	0.0001	0.0082	0.0171	0.0114
IBM_image_run_Mnozero11	Visual	0.0008	0	0.0045	0.0057	0.0095

Table 3. Image based retrieval results.

5 Compound Figure Separation

5.1 Method

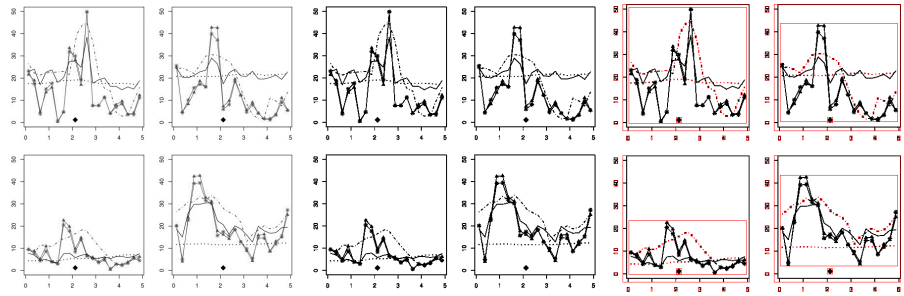
The method used for segmentation involves a combination of two approaches. The first approach uses an analysis of connected components in a binarized image, while the second approach, uses the common notation of subfigures using text.

In the first approach, the image is converted to grayscale, then binarized with a fixed threshold of 240 and analyzed for connected components. Very small components (dots, noise specks) are filtered out. Components that are contained within other are also joined to the containing components. The next step is to analyze the large components (those with bounding rectangle area greater than 500 pixels). This step uses an equivalence class to group components that are similar with regards to area and aspect ratio. The group containing the vast

majority of area and active pixels (non-background) is considered the group containing the sub-figures. Each member of this group is a sub-figure. If no such group is found, the method is considered failed, and the second approach is tried. Post processing of the sub-figure include classifying any remaining component by checking its distance to the nearest sub-figure bounding rectangle edge.

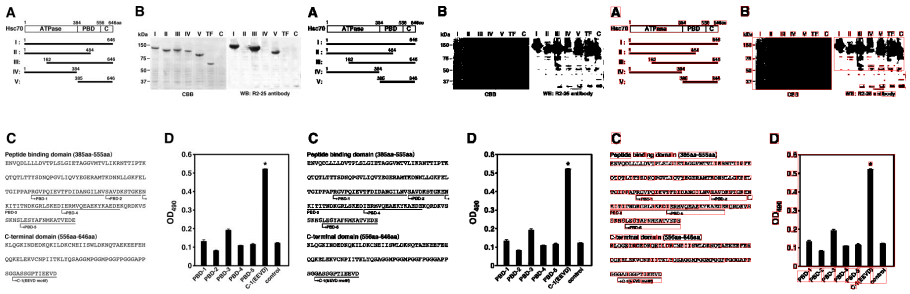
The second approach basically uses OCR to recognize all isolated components as letters.

Results with high enough confidence are considered letters and then equivalence is determined by letter component size. The actual sub-figure letters are selected by looking for a sequence of letters that are composed of the consecutive A,B,C,... which are also relatively arranged in a grid of some width and height (e.g. 2x2, 3x2, 2x4). This grid structure, if found, also directly dictates the bounds of the sub-figures. An example of the compound figure separation pipeline for each approach is shown in Figures 6 and 7.



(a) (b) (c)

Fig. 6. Compound image seaparation pipeline for the first approach: (a) input image, (b) binarization result and (c) identified connected components.



(a) (b) (c)

Fig. 7. Compound image seaparation pipeline for the second approach: (a) input image, (b) binarization result and (c) identified connected components.

5.2 Results

Even though we did not submit an official run for the task, we obtained the evaluation code from the organizers of the task and tested the performance of our algorithm on the official test set. Table 4 shows the results, compared to the official runs submitted by other teams. Our approach was purely visual and performed reasonably well. We are currently investigating further improvements to boost the performance of the algorithm.

Runs	Group name	Run type	Correctly classified in %
HESSO_CFS	medGIFT	Visual	84.64
nlm_multipanel_separation	ITI	Mixed	69.27
fcse-final-noempty	FINKI		68.59
IBM_compound_separation	IBM	Visual	**61.66**
HESSO_REGIONDETECTOR SCALE50_STANDARD	medGIFT	Visual	46.82

Table 4. Compound Figure Separation results.

6 Conclusions

In conclusion, for Modality Classfication we found that the textual and visual information provide complementary information. For the text based runs, Modality Taylored Keywords to work best. The best individual visual descriptor proved to be the Medical Semantic Model Vector. The combination not only of multiple descriptors, but also of different modeling strategies proved to be beneficial. We believe that more sophisticated multimodal fusion techniques could lead to further significant improvement over what was obtained in the submitted mixed runs.

For Case-Based retrieval, the use of topic modeling and the combination of mixed ontologies proved to be effective for the top query hits, as evidenced by the precision at 10 of our submitted runs. We also found that the number of topics employed during the training process produces a trade-off in performance.

Finally, experiments with two purely visual compound figure segmentation approaches, one based on connected component analysis and one based on subfigure text indexes detection, showed promise and we plan to further refine them in future iterations of this task.

References

1. Ahonen, T., Hadid, A., Pietikainen, M.: Face recognition with local binary patterns. In: ECCV. (2004) 469–481
2. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60, 2 (2004) 91–110

3. van Gemert, J.C., Veenman, C.J., Smeulders, A.W., Geusebroek, J.M.: Visual word ambiguity. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32**(7) (2010) 1271–1283
4. Van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32**(9) (2010) 1582–1596
5. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms (2008)
6. Merler, M., Huang, B., Xie, L., Hua, G., Natsev, A.: Semantic model vectors for complex video event recognition. Multimedia, IEEE Transactions on **14**(1) (feb. 2012) 88 –101
7. Fellbaum, C.: Wordnet: An electronic lexical database. MIT Press (1998)
8. Wu, S., Liu, H., Li, D., Tao, C., Musen, M., Chute, C., Shah, N.: Umls term occurrences in clinical notes: A large-scale corpus analysis. Journal of American Medical Informatics Association
9. Pedersen, T., Patwardhan, S., Michelizzi, J.: Word-net::similarity -measuring the relatedness of concepts. In: Fifth Annual Meeting of the North Ameri-can Chapter of the ACL (NAACL-04). (2004)
10. Garla, V., Re, V.L., Dorey-Stein, Z., Kidwai, F., Scotch, M., Womack, J., Justice, A., Brandt, C.: The yale ctakes extensions for document classification: architecture and application. Journal of the American Medical Informatics Association **18**(5) (2011) 614–620
11. Yan, R., Tesic, J., Smith, J.R.: Model-shared subspace boosting for multi-label classification. In: KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. (2007) 834–843
12. Cao, L., Chang, Y.C., Codella, N.C.F., Merler, M., Nguyen, Q.B., Smith, J.R.: Ibm t.j. watson research center, multimedia analytics: Modality classification and case-based retrieval tasks of imageclef2012. In: CLEF (Online Working Notes/Labs/Workshop). (2012)
13. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. IEEE Trans. Pattern Anal. Mach. Intell. **34**(3) (2012) 480–492
14. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. Journal of Machine Learning Research **9** (2008) 1871–1874
15. Cao, L., Luo, J., Liang, F., Huang, T.S.: Heterogeneous feature machines for visual recognition. In: ICCV. (2009) 1095–1102
16. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Computation **18**(7) (2006) 1527–1554
17. Kumari, D.M.U.R.G.P.: A study of meta-learning in ensemble based classifier. Engineering Science and Technology: An International Journal (ESTIJ) **2**(1) (February 2012) 36–41
18. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. In: SIGKDD Explorations. Volume 11. (2009)
19. M, S., T., G.: Probabilistic topic models. Lawrence Erlbaum **427** (2007)
20. McCallum, A.K.: Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu (2002)
21. DM, B., Andrew, N., Michael, J.: Latent dirichlet allocation. Journal of Machine Learning Research