Overview of the ImageCLEF 2012 medical image retrieval and classification tasks

Henning Müller[1,2], Alba G. Seco de Herrera [1], Jayashree Kalpathy–Cramer[3], Dina Demner Fushman[4], Sameer Antani[4], Ivan Eggel [1]

[1]University of Applied Sciences Western Switzerland, Sierre, Switzerland
[2]Medical Informatics, University of Geneva, Switzerland
[3]Harvard University, Cambridge, MA, USA
[4]National Library of Medicine (NLM), USA
henning.mueller@hevs.ch

Abstract. The ninth edition of the ImageCLEF medical image retrieval and classification tasks was organized in 2012. A subset of the open access collection of PubMed Central was used as the database in 2012, using a larger number of over 300'000 images than in 2011. As in previous years, there were three subtasks: modality classification, image–based and case–based retrieval.
A new hierarchy for article figures was created for the modality classification task. The modality detection could be one of the most important filters to limit the search and focus the results sets. The goal of the image–based and the case–based retrieval tasks were similar compared to 2011 adding mainly complexity.
The number of groups submitting runs has remained stable at 17, with the number of submitted runs remaining roughly the same with 202 (207 in 2011). Of these, 122 were image–based retrieval runs, 37 were case–based runs while the remaining 43 were modality classification runs. Depending on the exact nature of the task, visual, textual or multimodal approaches performed better.

1 Introduction

The CLEF 2012[1] labs continue the CLEF tradition of community–based benchmarking and complement it with workshops on emerging topics on information retrieval evaluation methodologies. Following the format introduced in 2010, two forms of labs were offered: labs could either be run as benchmarking activities campaign–style during the ten month period preceding the conference, or as workshop–style labs that explore possible benchmarking activities and provide a means to discuss information retrieval evaluation challenges from various perspectives.

ImageCLEF[2] [1–4] is part of CLEF and focuses on cross–language and language–independent annotation and retrieval of images. ImageCLEF has been organized since 2003. Four tasks were offered in 2012:

[1] http://www.clef2012.org/
[2] http://www.imageclef.org/

– medical image classification and retrieval;
– photo annotation and retrieval (large–scale web, Flickr, and personal photo tasks);
– plant identification;
– robot vision.

The medical image classification and retrieval task in 2012 is a use case of the PROMISE[3] network of excellence and is supported by the project. This task covers image modality classification and image retrieval with visual, semantic and mixed topics in several languages using a data collection from the biomedical literature. This year, there are three types of tasks in the medical image classification and retrieval task:

– modality classification;
– image–based retrieval;
– case–based retrieval.

This article presents the main results of the tasks and compares results between the various participating groups and the techniques employed.

2 Participation, Data Sets, Tasks, Ground Truth

This section describes the details concerning the set–up and the participation in the medical retrieval task in 2012.

2.1 Participation

In total over 60 groups registered for the medical tasks and obtained access to the data sets. ImageCLEF in total had over 200 registrations in 2012, with a bit more than 30% of the groups submitting results. 17 of the registered groups submitted results to the medical tasks, the same number as in previous years. The following groups submitted at least one run:

– Bioingenium (National University of Colombia, Colombia)*;
– BUAA AUDR (BeiHang University, Beijing, China);
– DEMIR (Dokuz Eylul University, Turkey);
– ETFBL (Faculty of Electrical Engineering Banja Luka, Bosnia and Herzegovina)*;
– FINKI (University in Skopje, Macedonia)*;
– GEIAL (General Electric Industrial Automation Limited, United States)*;
– IBM Multimedia Analytics (United States)*;
– IPL (Athens University of Economics and Business, Greece);
– ITI (Image and Text Integration Project, NLM, United States)*;
– LABERINTO (Universidad de Huelva, Spain);
– lambdasfsu (San Francisco State University, United States)*;

[3] http://www.promise-noe.eu/

– medGIFT (University of Applied Sciences Western Switzerland, Switzerland);
– MIRACL (Higher Institute of Computer Science and Multimedia of Sfax, Tunisia)*;
– MRIM (Laboratoire d'Informatique de Grenoble, France);
– ReDCAD (National School of Engineering of Sfax, Tunisia)*;
– UESTC (University of Electronic Science and Technology, China);
– UNED–UV (Universidad Nacional de Educacion a Distancia and Universitat de València, Spain);

Participants marked with a star had not participated in the medical retrieval task in 2011.

A total of 202 valid runs were submitted, 43 of which were submitted for modality detection, 122 for the image–based topics and 37 for the case–based topics. The number of runs per group was limited to ten per subtask and case–based and image–based topics were seen as separate subtasks in this view.

2.2 Datasets

In ImageCLEFmed 2012, a larger database than 2011 was provided using the same types of images and the same journals. The database contains over 300,000 images of 75'000 articles of the biomedical open access literature that allow free redistribution of the data. The ImageCLEF database is a subset of the PubMed Central[4] database containing in total over 1.5 million images. PubMedCentral contains all articles in PubMed that are open access but the exact copyright for redistribution varies among the journals.

2.3 Modality Classification

Previous studies [5, 6] have shown that imaging modality is an important information on the image for medical retrieval. In user–studies [7], clinicians have indicated that modality is one of the most important filters that they would like to be able to limit their search by. Many image retrieval websites (Goldminer, Yottalook) allow users to limit the search results to a particular modality [8]. Using the modality information, the retrieval results can often be improved significantly [9].

An improved ad–hoc hierarchy with 31 classes in the sections compound or multipane images, diagnostic images and generic biomedical illustrations was created based on the existing data set [10]. The following hierarchy was used for the modality classification, more complex than the classes in ImageCLEF 2011.

The class codes with descriptions are the following ([Class code] Description):

– $[COMP]$ Compound or multipane images (1 category)
– $[Dxxx]$ Diagnostic images:
  • $[DRxx]$ Radiology (7 categories):

[4] `http://www.ncbi.nlm.nih.gov/pmc/`

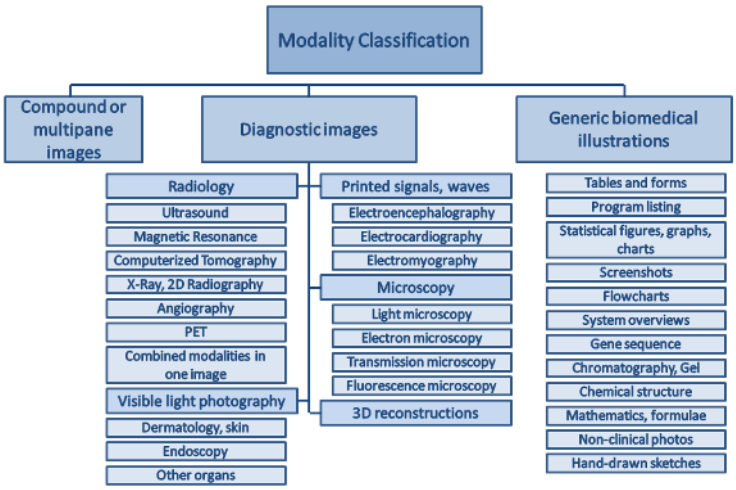

Fig. 1. The image classes hierarchy that was development for document images occurring in the biomedical open access literature.

  - $[DRUS]$ Ultrasound
  - $[DRMR]$ Magnetic Resonance
  - $[DRCT]$ Computerized Tomography
  - $[DRXR]$ X–Ray, 2D Radiography
  - $[DRAN]$ Angiography
  - $[DRPE]$ PET
  - $[DRCO]$ Combined modalities in one image
- $[DVxx]$ Visible light photography (3 categories):
  - $[DVDM]$ Dermatology, skin
  - $[DVEN]$ Endoscopy
  - $[DVOR]$ Other organs
- $[DSxx]$ Printed signals, waves (3 categories):
  - $[DSEE]$ Electroencephalography
  - $[DSEC]$ Electrocardiography
  - $[DSEM]$ Electromyography
- $[DMxx]$ Microscopy (4 categories):
  - $[DMLI]$ Light microscopy
  - $[DMEL]$ Electron microscopy
  - $[DMTR]$ Transmission microscopy
  - $[DMFL]$ Fluorescence microscopy
- $[D3DR]$ 3D reconstructions (1 category)
- $[Gxxx]$ Generic biomedical illustrations (12 categories):
  - $[GTAB]$ Tables and forms

- [$GPLI$] Program listing
- [$GFIG$] Statistical figures, graphs, charts
- [$GSCR$] Screenshots
- [$GFLO$] Flowcharts
- [$GSYS$] System overviews
- [$GGEN$] Gene sequence
- [$GGEL$] Chromatography, Gel
- [$GCHE$] Chemical structure
- [$GMAT$] Mathematics, formulae
- [$GNCP$] Non–clinical photos
- [$GHDR$] Hand–drawn sketches

For this hierarchy 1,000 training images and 1,000 test images were provided to the participants. Labels for the training images were known whereas labels for the test images were distributed after the results submission, only.

2.4 Image–Based Topics

The image–based retrieval task is the classic medical retrieval task, similar to the tasks organized from 2004 to 2011 where the query targets are single images. Participants were given a set of 22 textual queries (in English, Spanish, French and German) with 1–7 sample images for each query. The queries were classified into textual, mixed and semantic queries, based on the methods that are expected to yield the best results.

The topics for the image–based retrieval task were based on a selection of queries from search logs of the Goldminer radiology image search system [11]. Only queries occurring 10 times or more (about 200 queries) were considered as candidate topics for this task. A radiologist assessed the importance of the candidate topics, resulting in 50 candidate topics that were checked for at least occurring a few times in the database. The resulting 22 queries were then distributed among the participants and example query images were selected from a past collection of ImageCLEF [12].

2.5 Case–Based Topics

The case–based retrieval task was first introduced in 2009. This is a more complex task but one that we believe is closer to the clinical workflow. In this task, 30 case descriptions with patient demographics, limited symptoms and test results including imaging studies were provided (but not the final diagnosis). The goal was to retrieve cases including images that a physician would judge as relevant for differential diagnosis. Unlike the ad–hoc task, the unit of retrieval here was a case, not an image. The topics were created form an existing medical case database. Topics included a narrative text and several images.

2.6 Relevance Judgements

The relevance judgements were performed with the same on–line system as in 2008–2011 for the image–based topics as well as case–based topics. For the case–based topics, the system displays the article title and several images appearing in the text (currently the first six, but this can be configured). Judges were provided with a protocol for the process with specific details on what should be regarded as relevant versus non–relevant. A ternary judgement scheme was used again, wherein each image in each pool was judged to be "relevant", "partly relevant", or "non–relevant". Images clearly corresponding to all criteria were judged as "relevant", images for which relevance could not be accurately confirmed were marked as "partly relevant" and images for which one or more criteria of the topic were not met were marked as "non–relevant". Judges were instructed in these criteria and results were manually verified during the judgement process. As in previous years, judges were recruited by sending out an email to current and former students at OHSU's (Oregon Health and Science University) Department of Medical Informatics and Clinical Epidemiology. Judges, primarily clinicians, were paid a small stipend for their services. Many topics were judged by two or more judges to explore inter–rater agreements and its effects on the robustness of the rankings of the systems.

3 Results

This section describes the results of ImageCLEF 2012. Runs are ordered based on the tasks (modality classification, image–based and case–based retrieval) and the techniques used (visual, textual, mixed).

17 teams submitted at least one run in 2012, the same number than in 2011.

3.1 Modality Classification Results

The results of the modality classification task are compared using classification accuracy. With a higher number of classes, this task was more complex than in previous years. As seen in Table 1, the best result were obtained by the IBM Multimedia Analytics [13] group using visual methods (69.6%). In previous years combining visual and textual methods most often provided the best results. The best run using visual methods had a slightly better accuracy than the best run using mixed methods (66.2%) by the medGIFT group [14]. Only a single group submitted text–based results that performed worse than the average of all runs. The best run using textual methods alone obtained a much lower accuracy (41.3%).

Techniques Used for Visual Classification The IBM Multimedia Analytics team used multiple features extracted from a set of image granularities with Kernel approximation fusion in the best run [13]. A variety of image processing techniques were explored by the other participants. Multiple features were

Table 1. Results of the runs of modality classification task.

Run	Group	Run Type	Accuracy
medgift–nb–mixed–reci–14–mc	medGIFT	Mixed	66,2
medgift–orig–mixed–reci–7–mc	medGIFT	Mixed	64,6
medgift–nb–mixed–reci–7–mc	medGIFT	Mixed	63,6
Visual_Text_Hierarchy_w_Postprocessing_4_Illustration	ITI	Mixed	63,2
Visual_Text_Flat_w_Postprocessing_4_Illustration	ITI	Mixed	61,7
Visual_Text_Hierarchy	ITI	Mixed	60,1
Visual_Text_Flat	ITI	Mixed	59,1
medgift–b–mixed–reci–7–mc	medGIFT	Mixed	58,8
Image_Text_Hierarchy_Entire_set	ITI	Mixed	44,2
IPL_MODALITY_SVM_LSA_BHIST_324segs_50k_WithTextV	IPL	Mixed	23,8
Text_only_Hierarchy	ITI	Textual	41,3
Text_only_Flat	ITI	Textual	39,4
preds_Mic_Combo100Early_MAX_extended100	IBM Multimedia Analytics	Visual	69,6
LL_fusion_nfea_20_rescale	IBM Multimedia Analytics	Visual	61,8
preds_Mic_comboEarly_regular	IBM Multimedia Analytics	Visual	57,9
UESTC–MKL3	UESTC	Visual	57,8
UESTC–MKL2	UESTC	Visual	56,6
UESTC–MKL5	UESTC	Visual	55,9
UESTC–MKL6	UESTC	Visual	55,9
NCFC_ORIG_2_EXTERNAL_SUBMIT	IBM Multimedia Analytics	Visual	52,7
UESTC–SIFT	UESTC	Visual	52,7
Visual_only_Hierarchy	ITI	Visual	51,6
Visual_only_Flat	ITI	Visual	50,3
gist84_01_ETFBL	ETFBL	Visual	48,5
gist84_02_ETFBL	ETFBL	Visual	47,9
LL_2_EXTERNAL	IBM Multimedia Analytics	Visual	46,5
medgift–nb–visual–mnz–14–mc	medGIFT	Visual	42,2
medgift–nb–visual–mnz–7–mc	medGIFT	Visual	41,8
modality_visualonly	GEIAL	Visual	39,5
medgift–orig–visual–mnz–7–mc	medGIFT	Visual	38,1
medgift–b–visual–mnz–7–mc	medGIFT	Visual	34,2
NCFC_500_2_EXTERNAL_SUBMIT	IBM Multimedia Analytics	Visual	33,4
preds_Mic_comboLate_MAX_regular	IBM Multimedia Analytics	Visual	27,5
IPL_AllFigs_MODALITY_SVM_LSA_BHIST_324segs_50k	IPL	Visual	26,6
IPL_MODALITY_SVM_LSA_BHIST_324segs_50k	IPL	Visual	26,4
preds_Mic_comboLate_MAX_extended100	IBM Multimedia Analytics	Visua	22,1
UNED_UV_04_CLASS_IMG_ADAPTATIVEADJUST	UNED–UV	Visual	15,7
UNED_UV_03_CLASS_IMG_ADJUST2MINRELEVANTS	UNED–UV	Visual	13,4
UNED_UV_02_CLASS_IMG_ADJUST2AVGRELEVANTS	UNED–UV	Visual	13,1
UNED_UV_01_CLASS_IMG_NOTADJUST	UNED–UV	Visual	11,9
baseline–sift–k11–mc	medGIFT	Visual	11,1
testimagelabelres	GEIAL	Visual	10,1
ModalityClassificaiotnSubmit	BUAA AUDR	Manual	3,0

extracted from the images, most frequently scale–invariant feature transform (SIFT) variants [13–17], GIST (gist is not an acronym) [13], local binary patterns (LBP) [13, 17], edge and color histograms [13, 16–19] and gray value histograms [16]. Several texture features were also explored such us Tamura [13, 16–18], Gabor filters [16–18], Curvelets [13], a granulometric distribution function [19] and spatial size distribution [19]. For recognizing compound images ITI used an algorithm that detects sub–figure labels and the border of each sub–figure within a compound image [17].

k–Nearest Neighbors (kNN) [14], a logistic regression model [19] or multi–class support vector machines (SVMs) [13, 16–18] were employed to classify the images into the 32 categories. Only one group used hierarchical classification [17].

Three groups augmented the training data with additional examples for the categories [13, 14, 18]. Not all details of the training data expansion are clear and it needs to be assured that purely visual runs such as the best–performing run only use visual features for the training data set expansion.

Techniques Used for Classification Based on Text ITI [17] was the only group submitting a run for the textual modality classification task. They extracted the unified medical language system (UMLS) synonyms using the Essie system [20] and used it for term expansion when indexing enriched citations with Lucene/SOLR[5].

Techniques Used for Multimodal Classification Three groups submitted multimodal runs for the classification task. The medGIFT team obtained the best results [14] (66.2%). The approach fuses Bag–of–Visual–Words (BoVW) features based on SIFT and Bag–of–Colors (BoC) representing local image colors using reciprocal rank fusion. All three groups used techniques based on the Lucene search engine for the textual part and simple fusion techniques.

3.2 Image–Based Retrieval Results

13 teams submitted 36 visual, 54 textual and 32 mixed runs for the image–based retrieval task. The best result in terms of mean average precision (MAP) was obtained by ITI [17] using multimodal methods. The second best run was a purely textual run submitted by Bioingenium [21]. As in previous years, visual approaches achieved much lower results than the textual and multimodal techniques.

Visual Retrieval 36 of the 122 submitted runs used purely visual techniques. As seen in Table 2, DEMIR [22] achieved the best MAP, 0.0101, performing explicit grade relevance feedback. The second best run ($MAP = 0.0092$) was achieved also by DEMIR without applying relevance feedback. They combined color and edge directivity (CEDD) using combSUM [17, 21–23]. Bioingenium [21]

[5] `http://lucene.apache.org/`

submitted the third best run ($MAP = 0.0073$). They used a spatial pyramid extension for the CEDD.

In addition to the techniques used in the modality classification task, participants used visual features such as visual MPEG–7 features [22, 24], scalable color [24] and brightness/texture directionality histograms (BTDH) [22, 23]. Other techniques used are fuzzy color and texture histograms (FCTH) [17, 22, 23] and color layout (CL) [22, 24]. To extract these features most participants used tools such as Rummager [22, 23] or LIRE (Lucene Image Retrieval Engine) [24].

Table 2. Results of the **visual** runs for the medical image retrieval task.

Run Name	Group	MAP	GM-MAP	bpref	P10	P30
RFBr23+91qsum(CEDD,FCTH,CLD)max2012	DEMIR	0,0101	0,0004	0,0193	0,0591	0,0439
IntgeretedCombsum(CEDD,FCTH,CLD)max	DEMIR	0,0092	0,0005	0,019	0,05	0,0424
unal	Bioingenium	0,0073	0,0003	0,0134	0,0636	0,05
FOmixedsum(CEDD,FCTH,CLD)max2012	DEMIR	0,0066	0,0003	0,0141	0,0318	0,0288
edCEDD&FCTH&CLDmax2012	DEMIR	0,0064	0,0003	0,0154	0,0409	0,0318
medgift-lf-boc-bovw-mnz-ib	medGIFT	0,0049	0,0003	0,0138	0,0364	0,0364
Combined_LateFusion_Fileterd_Merge	ITI	0,0046	0,0003	0,0107	0,0318	0,0379
FilterOutEDFCTHsum2012	DEMIR	0,0042	0,0004	0,0109	0,0409	0,0364
finki	FINKI	0,0041	0,0003	0,0105	0,0318	0,0364
EDCEDDSUMmed2012	DEMIR	0,004	0,0003	0,0091	0,0364	0,0409
medgift-lf-boc-bovw-reci-ib	medGIFT	0,004	0,0002	0,0103	0,0227	0,0318
edFCTHsum2012	DEMIR	0,0034	0,0003	0,01	0,0318	0,0318
medgift-ef-boc-bovw-mnz-ib	medGIFT	0,0033	0,0003	0,0133	0,0364	0,0333
UNAL	Bioingenium	0,0033	0,0003	0,011	0,0455	0,0364
EDCEDD&FCTHmax2012	DEMIR	0,0032	0,0003	0,0111	0,0227	0,0303
medgift-ef-boc-bovw-reci-ib	medGIFT	0,003	0,0001	0,01	0,0273	0,0227
IntgeretedCombsum(CEDD,FCTH)max	DEMIR	0,0027	0,0003	0,0099	0,0045	0,0212
edMPEG7CLDsum2012	DEMIR	0,0026	0,0002	0,0058	0,0318	0,0242
UNAL	Bioingenium	0,0024	0,0001	0,0113	0,0091	0,0045
medgift-lf-boc-bovw-mnz-ib	medGIFT	0,0022	0,0001	0,0062	0,0227	0,0318
IPL_AUEB_DataFusion_LSA_SC_CL_CSH_64seg_20k	IPL	0,0021	0,0001	0,0049	0,0273	0,0242
IPL_AUEB_DataFusion_EH_LSA_SC_CL_CSH_64seg_100k	IPL	0,0018	0,0001	0,0053	0,0364	0,0258
IPL_AUEB_DataFusion_EH_LSA_SC_CL_CSH_64seg_20k	IPl	0,0017	0,0001	0,0053	0,0227	0,0273
IPL_AUEB_DataFusion_LSA_SC_CL_CSH_64seg_100k	IPL	0,0017	0,0002	0,0046	0,0364	0,0212
baseline-sift-early-fusion-ib	medGIFT	0,0017	0	0,0058	0,0227	0,0318
baseline-sift-late-fusion	medGIFT	0,0016	0	0,0048	0,0273	0,0318
IPL_AUEB_DataFusion_EH_LSA_SC_CL_CSH_64seg_50k	IPL	0,0011	0,0001	0,004	0,0136	0,0136
IPL_AUEB_DataFusion_LSA_SC_CL_CSH_64seg_50k	IPL	0,0011	0,0001	0,0039	0,0091	0,0121
Combined_Selected_Fileterd_Merge	ITI	0,0009	0	0,0028	0,0227	0,0258
reg_cityblock	lambdasfsu	0,0007	0	0,0024	0,0227	0,0197
reg_diffusion	lambdasfsu	0,0007	0	0,0023	0,0182	0,0182
tfidf_of_pca_euclidean	lambdasfsu	0,0005	0	0,0011	0,0136	0,0136
tfidf_of_pca_cosine	lambdasfsu	0,0005	0	0,0013	0,0091	0,0167
tfidf_of_pca_correlation	lambdasfsu	0,0005	0	0,0011	0,0136	0,0167
itml_cityblock	lambdasfsu	0,0001	0	0,0014	0,0045	0,003
itml_diffusion	lambdasfsu	0,0001	0	0,0015	0,0045	0,003

Textual Retrieval Table 3 shows that the Bioingenium [21] team achieved the best MAP using textual techniques (0.2182). They developed their own implementation of Okapi–BM25. The BUAA AUDR [18] team achieved the second best textual result (0.2081) with a run indexed with MeSH for query expansion

and modality prediction. The remaining participants explored a variety of retrieval techniques such as stop word and special character removal, tokenization and stemming (e. g. Porter stemmer) [19, 22–25]. For text indexing many groups used Terrier [22, 23, 26, 27]. In 2012, some groups included concept features [16, 25, 26] using tools such as MetaMap or MeSHUP. Query expansion [22, 28] was also explored.

Multimodal Retrieval The run with the highest MAP in the image retrieval task was a multimodal run submitted by the ITI team [17] (0.2377), see also Table 4. For this run various low–level visual descriptors were extracted to create the BoVW. This BoVW was combined with words taken from the topic description to form a multimodal query appropriate for Essie. ITI also submitted the second best mixed run ($MAP = 0.2166$) that has a slightly worse MAP than the best textual run ($MAP = 0.2182$).

Several late fusion strategies were used by the participants such as the product fusion algorithm [19], a linear weighed fusion strategy [23], reciprocal rank fusion [14], weighted combSUM [22] and combMNZ [14].

3.3 Case–based Retrieval Results

In 2012, 37 runs were submitted in the case–based retrieval task. As in previous years most of them were textual runs. Only the medGIFT team [14] submitted visual and multimodal case–based retrieval runs. Although textual runs achieved the best results, a mixed approach performs better than the average of all submitted runs in this task. Visual runs do not perform as well as most of the textual retrieval runs.

Visual Retrieval Table 5 shows the results using visual retrieval on the case–based task. The medGIFT team [14] is the only group that submitted a multimodal run in this task, using a combination of BoVW and BoC and obtaining the best accuracy in the multimodal classification task. The results also show that there can be an enormous difference combining the two base feature sets.

Textual Retrieval The medGIFT team [14] achieved the highest MAP, 0.169, among all submitted runs. For this run only the standard Lucene baseline was used. The second best run was submitted by MRIM ($MAP = 0.1508$) [28]. MRIM proposed a solution to the frequency shift thrugh a new counting strategy.

In addition to the techniques used in other tasks, the participants used semantic similarity [13, 16] measures. Moreover, three of the six groups participating used concept–based approaches [16, 25, 28]. The ITI team [17] used the Google Search API[6] to determine relevant disease names to correspond to signs and symptoms found in a topic case.

[6] `https://developers.google.com/custom-search/v1/overview`

Table 3. Results of the **textual** runs for the medical image retrieval task.

Run Name	Group	MAP	GM-MAP	bpref	P10	P30
UNAL	Bioingenium	0.2182	0.082	0.2173	0.3409	0.2045
AUDR_TFIDF_CAPTION[QE2]_AND_ARTICLE	BUAA AUDR	0.2081	0.0776	0.2134	0.3091	0.2045
AUDR_TFIDF_CAPTION[QE2]_AND_ARTICLE	BUAA AUDR	0.2016	0.0601	0.2049	0.3045	0.1939
IPL_A1T113C335M1	IPL	0.2001	0.0752	0.1944	0.2955	0.2091
IPL_A10T10C60M2	IPL	0.1999	0.0714	0.1954	0.3136	0.2076
TF_IDF	DEMIR	0.1905	0.0531	0.1822	0.3318	0.2152
AUDR_TFIDF_CAPTION_AND_ARTICLE	BUAA AUDR	0.1891	0.0508	0.1975	0.3318	0.1939
IPL_T10C60M2	IPL	0.188	0.0694	0.1957	0.3364	0.2076
AUDR_TFIDF_CAPTION[QE2]	BUAA AUDR	0.1877	0.0519	0.1997	0.3	0.2045
TF_IDF	DEMIR	0.1865	0.0502	0.1981	0.25	0.1515
Laberinto_MSH_PESO_2	Laberinto	0.1859	0.0537	0.1939	0.3318	0.1894
IPL_TCM	IPL	0.1853	0.0755	0.1832	0.3091	0.2152
IPL_T113C335M1	IPl	0.1836	0.0706	0.1868	0.3318	0.2061
UNAL	Bioingenium	0.1832	0.0464	0.1822	0.2955	0.1939
TF_IDF	DEMIR	0.1819	0.0679	0.1921	0.2864	0.1909
TF_IDF	DEMIR	0.1814	0.0693	0.1829	0.2864	0.1894
UESTC-ad-tc	UESTC	0.1769	0.0614	0.1584	0.3	0.1621
finki	FINKI	0.1763	0.0498	0.1773	0.2909	0.1864
Laberinto_MSH_PESO_1	Laberinto	0.1707	0.0512	0.1712	0.3318	0.1894
finki	FINKI	0.1704	0.0472	0.1701	0.3091	0.1833
Laberinto_MMTx_MSH_PESO_2	Laberinto	0.168	0.0555	0.1711	0.3227	0.1909
Terrier_CapTitAbs_BM25b0.75	ReDCAD	0.1678	0.0661	0.1782	0.2818	0.1712
Laberinto_MMTx_MSH_PESO_1	Laberinto	0.1677	0.0554	0.1701	0.3182	0.1879
AUDR_TFIDF_CAPTION[QE2]	BUAA AUDR	0.1673	0.037	0.1696	0.2955	0.1894
Laberinto_BL	Laberinto	0.1658	0.0477	0.1667	0.3	0.1939
AUDR_TFIDF_CAPTION	BUAA AUDR	0.1651	0.0467	0.1743	0.3	0.2076
AUDR_TFIDF_CAPTION	BUAA AUDR	0.1648	0.0441	0.1717	0.3318	0.1909
finki	FINKI	0.1638	0.0444	0.1644	0.3	0.1818
IPL_ATCM	IPL	0.1616	0.0615	0.1576	0.2773	0.1742
Laberinto_BL_MSH	Laberinto	0.1613	0.0462	0.1812	0.2682	0.1864
LIG_MRIM_IB_TFIDF_W_avdl_DintQ	MRIM	0.1586	0.0465	0.1596	0.3455	0.2136
HES-SO-VS_CAPTIONS_LUCENE	medGIFT	0.1562	0.0424	0.167	0.3273	0.1864
TF_IDF	DEMIR	0.1447	0.0313	0.1445	0.2864	0.1742
UESTC-ad-c	UESTC	0.1443	0.0352	0.1446	0.2409	0.1485
UESTC-ad-tcm	UESTC	0.1434	0.051	0.1397	0.2182	0.153
LIG_MRIM_IB_FUSION_TFIDF_W_TB_C_avdl_DintQ	MRIM	0.1432	0.0462	0.1412	0.2682	0.1955
LIG_MRIM_IB_FUSION_JM01_W_TB_C	MRIM	0.1425	0.0476	0.1526	0.2636	0.1924
HES-SO-VS_FULLTEXT_LUCENE	medGIFT	0.1397	0.0436	0.1565	0.2227	0.1379
LIG_MRIM_IB_TB_PIVv2_C	MRIM	0.1383	0.0405	0.1463	0.2864	0.1803
TF_IDF	DEMIR	0.1372	0.0466	0.1683	0.3	0.1818
Laberinto_MMTx_MSH	Laberinto	0.1361	0.0438	0.157	0.2091	0.1758
LIG_MRIM_IB_TFIDF_C_avdl_DintQ	MRIM	0.1345	0.0402	0.1304	0.2545	0.1682
LIG_MRIM_IB_TB_JM01_C	MRIM	0.1342	0.0396	0.142	0.2818	0.1652
LIG_MRIM_IB_TB_BM25_C	MRIM	0.1165	0.036	0.1276	0.2	0.1515
LIG_MRIM_IB_TB_TFIDF_C_avdl	MRIM	0.1081	0.0332	0.1052	0.1818	0.1167
UESTC-ad-cm	UESTC	0.106	0.0206	0.1154	0.2091	0.1379
UESTC_ad_tcm_mc	UESTC	0.101	0.0132	0.1223	0.2	0.1333
LIG_MRIM_IB_TB_DIR_C	MRIM	0.0993	0.0281	0.1046	0.1864	0.1379
AUDR_TFIDF_CAPTION_AND_ARTICLE	BUAA AUDR	0.0959	0.0164	0.1075	0.1636	0.1152
LIG_MRIM_IB_TB_TFIDF_C	MRIM	0.09	0.026	0.0889	0.1409	0.1136
UESTC-ad-cm-mc	UESTC	0.0653	0.0078	0.0846	0.1727	0.103
UNED_UV_01_TXT_AUTO_EN	UNED–UV	0.0039	0.0001	0.0055	0.0091	0.0076
UNAL	Bioingenium	0.0024	0.0001	0.0113	0.0091	0.0045

Table 4. Results of the **multimodal** runs for the medical image retrieval task.

Run Name	Group	MAP	GM-MAP	bpref	P10	P30
nlm-se	ITI	0.2377	0.0665	0.2542	0.3682	0.2712
Merge_RankToScore_weighted	ITI	0.2166	0.0616	0.2198	0.3682	0.2409
mixedsum(CEDD,FCTH,CLD)+1.7TFIDFmax2012	DEMIR	0.2111	0.0645	0.2241	0.3636	0.2242
mixedFCTH+1.7TFIDFsum2012	DEMIR	0.2085	0.0621	0.2204	0.3545	0.2152
medgift-ef-mixed-mnz-ib	medGIFT	0.2005	0.0917	0.1947	0.3091	0.2
mixedCEDD+1.7TFIDFsum2012	DEMIR	0.1954	0.0566	0.2096	0.3455	0.2182
nlm-lc	ITI	0.1941	0.0584	0.1871	0.2727	0.197
nlm-lc-cw-mf	ITI	0.1938	0.0413	0.1924	0.2636	0.2061
nlm-lc-scw-mf	ITI	0.1927	0.0395	0.194	0.2636	0.203
nlm-se-scw-mf	ITI	0.1914	0.0206	0.2062	0.2864	0.2076
Txt_Img_Wighted_Merge	ITI	0.1846	0.0538	0.2039	0.3091	0.2621
mixedsum(CEDD,FCTH,CLD)+TFIDFmax2012	DEMIR	0.1817	0.0574	0.1997	0.3409	0.2121
mixedFCTH+TFIDFsum2012	DEMIR	0.1816	0.0527	0.1912	0.3409	0.2076
finki	FINKI	0.1794	0.049	0.1851	0.3	0.1894
finki	FINKI	0.1784	0.0487	0.1825	0.2955	0.1864
nlm-se-cw-mf	ITI	0.1774	0.0141	0.1868	0.2909	0.2091
mixedCEDD+textsum2012	DEMIR	0.1682	0.0478	0.1825	0.3136	0.2061
FOmixedsum(CEDD,FCTH,CLD)+1.7TFIDFmax2012	DEMIR	0.1637	0.0349	0.1705	0.2773	0.1758
RFBr24+91qsum(CEDD,FCTH,CLD)+1.7TFIDFmax2012	DEMIR	0.1589	0.0424	0.1773	0.3136	0.1985
medgift-ef-mixed-reci-ib	medGIFT	0.1167	0.0383	0.1238	0.1864	0.1485
UNED_UV_04_TXTIMG_AUTO_LOWLEVEL_FEAT...	UNED–UV	0.004	0.0001	0.0104	0.0409	0.0258
UNED_UV_05_IMG_EXPANDED_FEATURES_UNIQUE...	UNED–UV	0.0036	0.0001	0.0111	0.0455	0.0303
UNED_UV_02_IMG_AUTO_LOWLEVEL_FEATURES	UNED–UV	0.0034	0.0001	0.0114	0.0455	0.0273
UNED_UV_08_IMG_AUTO_CONCEPTUAL_FEATURES	UNED–UV	0.0033	0.0001	0.0104	0.0227	0.0197
IPL_AUEB_SVM_CLASS_LSA_BlockHist324Seg_50k	IPL	0.0032	0.0002	0.0103	0.0409	0.0303
IPL_AUEB_SVM_CLASS_TEXT_LSA_BlockHist324Seg_50k	IPL	0.0025	0.0001	0.0095	0.0318	0.0258
IPL_AUEB_CLASS_LSA_BlockColorLayout64Seg_50k	IPL	0.0023	0.0002	0.0095	0.0318	0.0227
UNED_UV_09_TXTIMG_AUTO_CONCEPTUAL_FEAT...	UNED–UV	0.0021	0.0001	0.005	0.0091	0.0061
IPL_AUEB_CLASS_LSA_BlockColorLayout64Seg_20k	IPL	0.0019	0.0001	0.0066	0.0227	0.0197
UNED_UV_03_TXTIMG_AUTO_LOWLEVEL_FEAT...	UNED–UV	0.0015	0.0001	0.0037	0.0045	0.0061
UNED_UV_07_TXTIMG_AUTO_EXPANDED_FEAT...	UNED–UV	0.0015	0.0001	0.0036	0.0045	0.0061
UNED_UV_06_TXTIMG_AUTO_EXPANDED_FEAT...	UNED–UV	0.0013	0.0001	0.0034	0.0091	0.0045

Table 5. Results of the **visual** runs for the medical case–based retrieval task.

Run Name	Group	MAP	GM-MAP	bpref	P10	P30
medgift-lf-boc-bovw-reci-IMAGES-cb	medGIFT	0,0366	0,0014	0,0347	0,0269	0,0141
medgift-lf-boc-bovw-mnz-IMAGES-cb	medGIFT	0,0302	0,001	0,0293	0,0231	0,009
baseline-sift-early-fusion-cb	medGIFT	0,0016	0	0,0032	0,0038	0,0013
baseline_sift_late_fusion_cb	medGIFT	0,0008	0	0	0,0038	0,0013
medgift-ef-boc-bovw-reci-IMAGES-cb	medGIFT	0,0008	0,0001	0,0007	0	0,0013
medgift-ef-boc-bovw-mnz-IMAGES-cb	medGIFT	0,0007	0	0	0	0,0013

Table 6. Results of the **textual** runs for the medical case–based retrieval task.

Run Name	Group	MAP	GM-MAP	bpref	P10	P30
HES-SO-VS_FULLTEXT_LUCENE	medGIFT	0,169	0,0374	0,1499	0,1885	0,109
LIG_MRIM_CB_FUSION_DIR_W_TA_TB_C	MRIM	0,1508	0,0322	0,1279	0,1538	0,1167
LIG_MRIM_CB_FUSION_JM07_W_TA_TB_C	MRIM	0,1384	0,0288	0,11	0,1615	0,1141
UESTC_case_f	UESTC	0,1288	0,025	0,1092	0,1231	0,0821
UESTC-case-fm	UESTC	0,1269	0,0257	0,1117	0,1231	0,0821
LIG_MRIM_CB_TFIDF_W_DintQ	MRIM	0,1036	0,0167	0,077	0,0846	0,0705
nlm-lc-total-sum	ITI	0,1035	0,0137	0,1053	0,1	0,0628
nlm-lc-total-max	ITI	0,1027	0,0125	0,1055	0,0923	0,0538
nlm-se-sum	ITI	0,0929	0,013	0,0738	0,0769	0,0667
nlm-se-max	ITI	0,0914	0,0128	0,0736	0,0769	0,0667
nlm-lc-sum	ITI	0,0909	0,0133	0,0933	0,1231	0,0654
LIG_MRIM_CB_TA_TB_JM07_C	MRIM	0,0908	0,0156	0,0799	0,1308	0,0744
LIG_MRIM_CB_TA_TB_BM25_C	MRIM	0,0895	0,0143	0,0864	0,1231	0,0654
LIG_MRIM_CB_TA_TB_DIR_C	MRIM	0,0893	0,0137	0,0804	0,1192	0,0692
LIG_MRIM_CB_TA_TB_PIVv2_C	MRIM	0,0865	0,0158	0,0727	0,1192	0,0795
nlm-lc-max	ITI	0,084	0,0109	0,0886	0,0923	0,0603
LIG_MRIM_CB_TA_TFIDF_C_DintQ	MRIM	0,0789	0,014	0,0672	0,0923	0,0692
nlm-se-frames-sum	ITI	0,0771	0,0052	0,0693	0,0692	0,0526
HES-SO-VS_CAPTIONS_LUCENE	medGIFT	0,0696	0,0028	0,0762	0,0962	0,0615
LIG_MRIM_CB_TA_TB_TFIDF_C_avdl	MRIM	0,0692	0,0127	0,0688	0,0769	0,0692
nlm-se-frames-max	ITI	0,0672	0,0031	0,0574	0,0538	0,05
LIG_MRIM_CB_TA_TB_TFIDF_C	MRIM	0,0646	0,0114	0,0624	0,0692	0,0641
ibm-case-based	IBM	0,0484	0,0023	0,0439	0,0577	0,0449
R1_MIRACL	MIRACL	0,0421	0,005	0,026	0,0538	0,0462
R4_MIRACL	MIRACL	0,0196	0,0008	0,0165	0,0308	0,0282
R3_MIRACL	MIRACL	0,012	0,0004	0,0087	0,0192	0,0218
R6_MIRACL	MIRACL	0,0111	0,0004	0,0074	0,0192	0,0128
R5_MIRACL	MIRACL	0,0024	0	0,0022	0,0038	0,0013
R2_MIRACL	MIRACL	0	0	0,0002	0	0

Multimodal Retrieval As in the visual case–based task, only the medGIFT team [14] submitted multimodal case–based runs. The runs combine the visual approach based on BoVW and BoC with a Lucene baseline and obtained averaged results when using the combMNZ fusion.

Table 7. Results of the **multimodal** runs for the medical case retrieval task.

Run Name	Group	MAP	GM-MAP	bpref	P10	P30
medgift-ef-mixed-mnz-cb	medGIFT	0,1017	0,0175	0,0857	0,1115	0,0679
medgift-ef-mixed-reci-cb	medGIFT	0,0514	0,009	0,0395	0,0654	0,0564

4 Conclusions

As in previous years, the largest number of runs submitted for the image–based retrieval task. However, in 2012 there were 122 runs in this task, eight less than in 2011. For the case–based retrieval task the number of runs also decreased to 37 (43 in 2011). On the other hand, the number submitted runs at the modality classification task increased to 43 (34 in 2011).

There are still different situations as to whether visual, textual or combined techniques perform better depending on the task. For the modality classification, a visual run achieved the best accuracy using training data extension. In the case of the image–based retrieval task, multimodal runs obtained best results. Finally, for the case–based retrieval task textual runs obtained the best results.

In 2011, the Xerox team [29] that did not participate in 2012 explored the expansion of the training set. This approach achieved the best accuracy for the modality classification task. In 2012, three teams applied expansion of the training set and also obtained good results. This evolution of techniques is a good example of the added value of evaluation campaigns such as ImageCLEF showing the improvements due to specific techniques.

Many groups explored the same or similar descriptors obtaining often quite differing results. This shows that particularly the tuning of existing techniques and the intelligent combination of results fusion can lead to optimal results. Often, the differences in techniques are quite small and more on intelligent feature combinations might be necessary to reach conclusive results.

5 Acknowledgements

We would like to thank the EU FP7 projects Khresmoi (257528), PROMISE (258191) and Chorus+ (249008) for their support as well as the Swiss national science foundation with the MANY project (number 205321–130046).

References

1. Müller, H., Clough, P., Deselaers, T., Caputo, B., eds.: ImageCLEF – Experimental Evaluation in Visual Information Retrieval. Volume 32 of The Springer International Series On Information Retrieval. Springer, Berlin Heidelberg (2010)
2. Clough, P., Müller, H., Sanderson, M.: The CLEF cross–language image retrieval track (ImageCLEF) 2004. In Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B., eds.: Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign. Volume 3491 of Lecture Notes in Computer Science (LNCS)., Bath, UK, Springer (2005) 597–613
3. Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: CLEF 2007 Proceedings. Volume 5152 of Lecture Notes in Computer Science (LNCS)., Budapest, Hungary, Springer (2008) 473–491
4. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., García Seco de Herrera, A., Tsikrika, T.: The CLEF 2011 medical image retrieval and classification tasks. In: Working Notes of CLEF 2011 (Cross Language Evaluation Forum). (September 2011)
5. Kalpathy-Cramera, J., Hersh, W.: Automatic image modality based classification and annotation to improve medical image retrieval. Studies in Health Technology and Informatics **129** (2007) 1334–1338
6. Csurka, G., Clinchant, S., Jacquet, G.: Medical image modality classification and retrieval. In: 9th International Workshop on Content-Based Multimedia Indexing, IEEE (2011) 193–198

7. Markonis, D., Holzer, M., Dung, S., Vargas, A., Langs, G., Kriewel, S., Müller, H.: A survey on visual information search behavior and requirements of radiologists. Methods of Information in Medicine (2012) Forthcoming
8. Zhang, D., Lu, G.: Review of shape representation and description techniques. Pattern Recognition (1) (2004) 1–19
9. Tirilly, P., Lu, K., Mu, X., Zhao, T., Cao, Y.: On modality classication and its use in text–based image retrieval in medical databases. In: Proceedings of the 9th International Workshop on Content–Based Multimedia Indenxing. CBMi2011 (2011)
10. Müller, H., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S.: Creating a classification of image types in the medical literature for visual categorization. In: SPIE medical imaging. (2012)
11. Tsikrika, T., Müller, H., Kahn Jr., C.E.: Log analysis to understand medical professionals' image searching behaviour. In: Proceedings of the 24th European Medical Informatics Conference. MIE2012 (2012)
12. Hersh, W., Müller, H., Kalpathy-Cramer, J., Kim, E., Zhou, X.: The consolidated ImageCLEFmed medical image retrieval task test collection. Journal of Digital Imaging **22**(6) (2009) 648–655
13. Cao, L., Chang, Y.C., Codella, N., Merler, M.: IBM t.j. watson research center, multimedia analytics: Modality classification and case–based retrieval task of ImageCLEF2012. In: Working Notes of CLEF 2012. (2012)
14. García Seco de Herrera, A., Markonis, D., Eggel, I., Müller, H.: The medGIFT group in ImageCLEFmed 2012. In: Working Notes of CLEF 2012. (2012)
15. Collins, J., Okada, K.: A comprative study of similarity measures for content–based medical image retrieval. In: Working Notes of CLEF 2012. (2012)
16. Wu, H., Sun, K., Deng, X., Zhang, Y., Che, B.: UESTC at ImageCLEF 2012 medical tasks. In: Working Notes of CLEF 2012. (2012)
17. Simpson, M.S., You, D., Rahman, M.M., Demmer-Fushman, D., Antani, S., Thoma, G.: ITI's participation in the ImageCLEF 2012 medical retrieval and classification tasks. In: Working Notes of CLEF 2012. (2012)
18. Song, W., Zhang, D., Luo, J.: BUAA AUDR at ImageCLEF 2012 medical retrieval task. In: Working Notes of CLEF 2012. (2012)
19. Castellanos, A., Benavent, J., Benavent, X., García-Serrano, A.: Using visual concep features in a multimodal retrieval system for the medical collection at ImageCLEF2012. In: Working Notes of CLEF 2012. (2012)
20. Ide, N.C., Loane, R.F., Demner-Fushman, D.: Application of information technology: Essie: A concept–based search engine for structured biomedical text. Journal of the American Medical Informatics Association **14**(3) (2007) 253–263
21. Vanegas, J.A., Caicedo, J.C., Camargo, J., Ramos, R., González, F.A.: Bioingenium at ImageCLEF: Textual and visual indexing for medical images. In: Working Notes of CLEF 2012. (2012)
22. Vahid, A.H., Alpkocak, A., Hamed, R.G., Caylan, N.M., Ozturkmenoglu, O.: DEMIR at ImageCLEFmed 2012: Inter–modality and intra–modality integrated combination retrieval. In: Working Notes of CLEF 2012. (2012)
23. Kitanovski, I., Dimitrovski, I., Loskovska, S.: FCSE at ImageCLEF 2012: Evaluating techniques for medical image retrieval. In: Working Notes of CLEF 2012. (2012)
24. Stathopoulos, S., Sakiotis, N., Kalamboukis, T.: IPL at CLEF 2012 medical retrieval task. In: Working Notes of CLEF 2012. (2012)

25. Majdoubi, J., Loukil, H., Tmar, M., Gargourri, F.: Medical case–based retrieval by using a language model: MIRACL at ImageCLEF 2012. In: Working Notes of CLEF 2012. (2012)
26. Gasmi, K., Torjmen-Khemakhem, M., Ben Jemaa, M.: Word indexing versus conceptual indexing in medical image retrieval (ReDCAD participation at ImageCLEF medical image retrieval 2012). In: Working Notes of CLEF 2012. (2012)
27. Crespo, M., Mata, J., Maña, M.J.: LABERINTO at ImageCLEF 2012 medical image retrieval tasks. In: Working Notes of CLEF 2012. (2012)
28. Abdulahhad, K., Chevallet, J.P., Berrut, C.: MRIM at ImageCLEF2012. from words to concepts: A new couting approach. In: Working Notes of CLEF 2012. (2012)
29. Csurka, G., Clinchant, S., Jacquet, G.: XRCE's participation at medical image modality classification and ad–hoc retrieval task of ImageCLEFmed 2011. In: Working Notes of CLEF 2011. (2011)