The final publication is available at

http://link.springer.com/chapter/10.1007/978-3-642-40802-1_26

ImageCLEF 2013: the vision, the data and the open challenges

Barbara Caputo[a], Henning Muller [b], Bart Thomee[c], Mauricio Villegas[d],
Roberto Paredes[d], David Zellhofer[e], Herve Goeau[f], Alexis Joly[g], Pierre
Bonnet[h], Jesus Martinez Gomez[i], Ismael Garcia Varea[i], and Miguel Cazorla[l].

[a] Idiap Research Institue, Martigny, Switzerland
[b] University of Applied Sciences Western Switzerland in Sierre, Switzerland
[c] Yahoo! Research, Barcelona, Spain
[d] ITI/DSIC, Universitat Politècnica de València, Spain
[e] Brandenburg University of Technology, Germany
[f] INRIA-IMEDIA, Paris, France
[g] INRIA-ZENITH, Montpellier, France
[h] CIRAD, UMR AMAP, Montpellier, France
[i] University of Castilla-La Mancha, Albacete, Spain
[l] University of Alicante, Alicante, Spain

Abstract. This paper presents an overview of the ImageCLEF 2013
lab. Since its first edition in 2003, ImageCLEF has become one of the
key initiatives promoting the benchmark evaluation of algorithms for the
cross-language annotation and retrieval of images in various domains,
such as public and personal images, to data acquired by mobile robot
platforms and botanic collections. Over the years, by providing new data
collections and challenging tasks to the community of interest, the Im-
ageCLEF lab has achieved an unique position in the multi lingual image
annotation and retrieval research landscape. The 2013 edition consisted
of three tasks: the photo annotation and retrieval task, the plant identi-
fication task and the robot vision task. Furthermore, the medical anno-
tation task, that traditionally has been under the ImageCLEF umbrella
and that this year celebrates its tenth anniversary, has been organized
in conjunction with AMIA for the first time. The paper describes the
tasks and the 2013 competition, giving an unifying perspective of the
present activities of the lab while discussion the future challenges and
opportunities.

1 Introduction

Since its first edition in 2003, the ImageCLEF lab initiative has focused on
providing an evaluation forum for the cross-language annotation and retrieval
of images [1]. The main motivation behind ImagCLEF is the need to support
multilingual users from a global community accessing the ever growing body
of visual information. Thus, the main goal of ImageCLEF is to support the
advancement of the field of visual media analysis, indexing, classification, and
retrieval, by developing the necessary infrastructure for the evaluation of visual

	2009	2010	2011	2012	2013
Total Participations	65	47	43	51	42
Total Registrations	84	112	141	209/106	219/95

Fig. 1. Number of registered groups versus number of groups that submitted at least one valid run since 2009. In 2012 and 2013, we report also the total number of groups that initiated the registration process but that, for several reasons, were not able to complete it in time.

information retrieval systems operating in monolingual, language-independent and multi-modal contexts, providing reusable resources for such benchmarking purposes.

To meet these objectives, ImageCLEF organises tasks that benchmark the annotation and retrieval of diverse images such as general photographic and medical images, as well as domain-specific tasks such as plant identification and robot vision. These evaluation tasks aim to support and promote research that addresses key challenges in the field including: 1) visual image annotation with concepts at various levels of abstraction that relies not only on manual, and thus reliable, training data but also on automatically acquired and thus noisy, labelled samples, 2) scientific multimedia data management through the particular case of botanical data identification, and 3) the shift in the area of robot vision from visual place recognition to multimodal place recognition. Moreover, the ImageCLEF 2013 lab has maintained its decade long traditional commitment to medical informatics by helping organizing a challenge on modality classification and retrieval in the medical domain. The aim is to move closer to clinical practice and routine through classification tasks that consider complex, hierarchically organised classes of modalities and retrieval tasks that support medical practitioners in their decision making. This challenge has moved for the first time in 2013 from ImageCLEF in conjunction with the American Medical Informatics Association (AMIA) annual symposium.

Over the years, ImageCLEF has had a significant influence on the visual information retrieval field by benchmarking various retrieval and annotation tasks and by making available the large and realistic test collections built in the context of its activities. Many research groups have participated over the years in its evaluation campaigns and even more have acquired its datasets for experimentation. Figure 1 shows the number of registered groups, and of groups that eventually submitted a run, since 2008. In 2013, over 200 research groups regis-

tered, with 42 of those submitting runs officially to the ImageCLEF tasks. The impact of ImageCLEF can also be seen by its significant scholarly impact indicated by the substantial numbers of its publications and their received citations [2].

The rest of the paper is organized as follows: section 2 describes the three subtasks of the 2013 edition: the photo annotation and retrieval task (section 2.1), the plant identification task (section 2.2), and the robot vision task (section 2.3). Section 2.4 describes the AMIA associated medical tasks. We conclude with an overall discussion, and pointing towards the challenges ahead and possible new directions for ImageCLEF 2014.

2 ImageCLEF 2013: the tasks, the data and participation

The 2013 edition of ImageCLEF consisted of three main tasks, plus one task associated with the AMIA 2013 meeting: the photo annotation and retrieval task, the plant identification task, the robot vision task and, jointly with AMIA, the medical task. These tasks had the goal to benchmark the annotation and retrieval of diverse images such as general photographic, as well as domain-specific tasks such as plant identification and robot vision. The overall aim is to support and promote research that addresses key challenges in the field including:

– visual image annotation with concepts at various levels of abstraction that relies not only on manual, and thus reliable, training data, but also on automatically acquired, and thus noisy, labelled samples,
– scientific multimedia data management through the particular case of botanical data identification, and
– the shift in the area of robot vision from visual place recognition to multimodal place recognition.

In the rest of the section, we give an overview account, for each task, of its historical perspective within ImageCLEF, of its 2013 objective and task, and of the task participation and relative results.

2.1 The photo annotation and retrieval task

Automatic concept detection within images is a challenging research problem, as of today yet unsolved. Despite considerable research efforts the so-called semantic gap has not yet been successfully breached, in terms of being able to detect semantic concepts within any kind of imagery for any kind of concept as accurately as real people can. ImageCLEF's photo annotation and retrieval task aims to advance the state of the art in multimedia research by acting as a platform to foster interaction and collaboration between researchers and by providing a realistic and challenging benchmark for visual concept detection, annotation and retrieval in the context of personal photo and web image collections.

Fig. 2. Exemplar images for the photo annotation task. The top row shows images obtained from a web search query of 'rainbow'; the bottom row shows images from a web search query of 'sun'.

Past Editions Annotation and retrieval of web images and personal photographs has been part of ImageCLEF since its very first edition in 2003. In the early years the focus was on retrieving relevant images from a web collection given (multilingual) queries, while from 2006 onwards annotation tasks were also held, initially aimed at object detection, but more recently also covering semantic concepts. Between 2009 and 2012 the photo annotation and retrieval tasks were based upon various subsets of the MIRFLICKR collection [3, 4], where every year the list of concepts to detect was updated in order to cover a wider selection of concept types, thus making the task more challenging. With the aim of providing new challenges to the research community, in 2012 two novel subtasks were introduced, one on annotation without requiring any manually labeled training data [5], and the other on retrieval in the context of personal photo collections [6]. These two paths have been continued for this year's task, and they are described in more details in the following.

Objective and Task for 2013 Edition This year's task has been divided into two separate subtasks, one entitled *Scalable Concept Image Annotation* and the other *Personal Photo Retrieval*. Each of the subtasks focuses on the two directions of research in this field on which the subtask organizers agreed that deserve more attention.

Annotation subtask: Image concept detection generally has relied on training data that has been manually, and thus reliably annotated, an expensive and laborious endeavor that cannot easily scale, particularly as the number of concepts grows. However, images for any topic can be cheaply gathered from the

Fig. 3. Exemplar images for the personal photo annotation task. The top row shows samples of the Visual Concept 'Asian Temple Interior'; the bottom row shows samples of the Event Class 'Rock Concert'.

web, along with associated text from the webpages that contain the images. The degree of relationship between these web images and the surrounding text varies greatly, i.e., the data is very noisy, but overall this data contains useful information that can be exploited to develop annotation systems. Likewise there are other resources available that can help to determine the relationships between text and semantic concepts, such as dictionaries or ontologies. The goal of this subtask was to evaluate different strategies to deal with the noisy data so that it can be reliably used for annotating images from practically any topic. Participants were provided with a training set composed of images and corresponding webpage text, and for the given development/test set they had to detect the corresponding concepts for each image using only the input image, the provided training set and any other automatically obtained resources.

Data The data used in this subtask is mostly the same as the one from last year's task [5], although there are differences [7]. The training set is composed of visual and textual features for 250,000 images downloaded from the web by querying popular search engines. The development and test sets have 1,000 and 2,000 images, respectively, which include only visual features and the corresponding hand labeled concepts ground truth. Figure 2 shows some exemplar images that illustrate the type of challenges addressed in the task. For further details, please refer to [7].

Personal photo retrieval subtask: This year's subtask has a focus on different retrieval usage scenarios and user groups. That is, the subtask reveals whether the tested algorithms are stable in terms of retrieval quality for different user groups. In order to associate relevance assessments with different user groups, the assessors had to answer a questionnaire (see [8]). The subtask is ad-hoc, i.e., no additional training data is released. The participants have to rely on multiple QBE documents and/or browsing data and are asked to find the best matching documents illustrating an event or depicting a visual concept. Thus, an

additional objective of this task is to find out whether the participating retrieval systems can exploit data from different search strategies, i.e., query-by-example and browsing data, in order to find both visual concepts and photos depicting events. To solve the task, the participants have access to pre-extracted visual low-level features, metadata, but are also free to use their own techniques.

Data The subtask uses the same document corpus as in 2012 [6], i.e., 5,555 images that have been sampled from 19 personal photo collections of layperson photographers. In contrast to the last year's pilot phase, the amount of queries has been increased and the queries are no longer subdivided into events and visual concepts. Additionally, the participants have access to a baseline system that can be used for feature extraction. Figure 3 shows some exemplar images that illustrate the type of challenges addressed in the task. More detailed information is available in a separate publication [8].

Participants and Results Generally speaking, the participation was excellent. In total, 18 groups took part in the task and submitted 84 runs, of which 26 runs were submitted by 7 groups to the retrieval subtask, whereas the remaining 58 runs were submitted by 13 groups to the annotation subtask. The following is a very brief summary of the results obtained for each subtask. For further details and analysis, the readers should refer to the corresponding overview paper, [7] or [8].

Annotation subtask results: In comparison to last year (the first edition of this subtask), this year's results have been much more interesting, even though the challenge has remained mostly the same. The main reason for this is the significantly greater number of participants and submissions. The participating groups have explored several interesting ideas to tackle the proposed problem, which gives hand to a more richer discussion. Figure 4 presents a graph that compares all of the submitted runs using the annotation mean F-measure (MF_1), measured both for the test samples and for the concepts. Most of the groups obtained a very impressive improvement in performance compared to the baselines. The most interesting aspect of the results was that even though one system outperformed the rest, many of the ideas proposed by the participants are complementary, so considerable improvements could be expected in future works. For further details, please refer to the subtask overview paper [7].

Personal photo retrieval subtask results: The best performing groups – ISI and DBIS – used visual low-level features and metadata to solve the task. While ISI used relevance feedback for all of their runs, DBIS used this technique only for run #3. In accordance with the findings of the last years' ImageCLEF tasks, there is evidence that the utilization of multiple modalities increases the retrieval effectiveness. Table 1 shows an excerpt of the average results in order to provide an overview over the general retrieval effectiveness achieved by the participants of the subtask. The user group-specific results are available at the subtask's
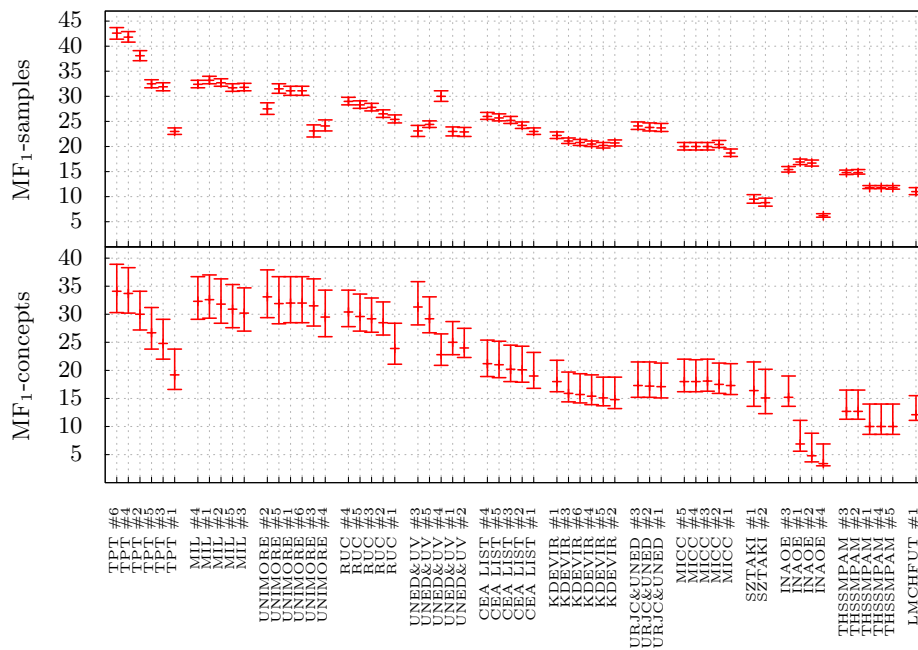
Fig. 4. Graphs showing the test set performance measures (in %) for all the submissions for the annotation subtask. The error bars correspond to the 95% confidence intervals computed using Wilson's method.

website[1]. Regarding the effectiveness variance over the different user groups, the results are not very clear. There are only minor differences between the user groups. For a discussion of this effect and a complete overview over the results, please refer to [8].

2.2 The plant identification task

If agricultural development is to be successful and biodiversity is to be conserved, then accurate knowledge of the identity, geographic distribution and uses of plants is essential. Unfortunately, such basic information is often only partially available for professional stakeholders, scientists and citizens. So that simply identifying plant species is usually a very difficult task, even for professionals. Using image retrieval technologies is nowadays considered by botanists as a promising direction in reducing this taxonomic gap. ImageCLEF plant identification task, funded by the French project Pl@ntNet and the EU coordination action CHORUS+, is aimed at evaluating recent advances of the multimedia IR community on this challenging problem.

[1] http://imageclef.org/2013/photo/retrieval#results
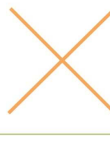
Table 1. Summary of the averaged results for the personal photo retrieval subtask (excerpt of the best submissions per group).

Group	Run ID	map_cut_100	ndcg_cut_10	ndcg_cut_20	ndcg_cut_30
DBIS	run3	0.3954	0.7197	0.6798	0.6546
FINKI	run2	0.1375	0.5510	0.4398	0.3881
IPL	IPL13_visual_r4	0.1162	0.5152	0.4173	0.3713
ISI	4	0.5034	0.2167	0.3132	0.3716
ThssMpam4	5000_TI_CR	0.070	0.4005	0.3051	0.2676
ThssMpam4	5000_TI_NCR	0.070	0.4009	0.3050	0.2675
VCTLab	2	0.0783	0.3574	0.3047	0.2754
WIDE_IO	WideIO	0.0584	0.3253	0.2501	0.2192

Past Editions Each year since 2011, the task is becoming closer to a real-world scenario thanks to the observations feed of a French social network specialized in botany (Tela Botanica). The underlying citizen science project aims at covering the entire French flora with a sufficiently rich and balanced collection of pictures. The dataset used for the 2013 campaign covered 250 species of herbs and trees living in France area (i.e. the most represented ones in the whole collected social data since 2011). Contrary to the two previous years that were exclusively focused on leaf images (of tree species only), the coverage of the 2013 task was extended to six different types of view of the plant: leaf scans (or scan-like), leaf photographs, flower photographs, fruit photographs, bark photographs, and the entire view of the plant. A separate evaluation score was computed for the two main categories of images, i.e. scans (or scan-like pictures) vs. photographs (with natural background). Proportions were around 42% of scans and scan-like pictures of leaves vs. 58% of photographs with a natural background (more precisely 16% of leaves, 18% of flowers, 8% of fruits, 8% of stems and 8% of entire). The whole database contained around 26k images collected by 327 distinct contributors, living in different regions in France, equipped with various cameras and at different periods of the year. This makes the task much more realistic than any previous data built for the evaluation of content-based plant identification methods.

Objective and Task for the 2013 Edition The precise goal of the 2013 task was to retrieve the correct species among the top k species of a ranked list of returned species, one list for each image of a test dataset. Participants received a first training set of annotated images in order to explore different techniques and train their system. Six weeks later participants received the test set containing images without species labels. Then participants were allowed to submit until 4 run files, most of the time related to variations of one same method. A particular attention was paid when splitting the data into training and test subsets to avoid any bias. Several pictures in the dataset might actually depict the same individual plant (or neighboring plants) observed in the same conditions (same person, day, device, lightening conditions, etc.). Randomly splitting images in

	Sheet As Background	Natural Background				
		Leaf	Flower	Fruit	Stem	Entire
Kaki Persimmon						
Boxelder maple						
Common mallow				✕	✕	
Common agrimony	✕					

Fig. 5. Examples of the different views used in the database: scan or scan-like images of leaves associated to a SheetAsBackground category, and photographs of leaves, flowers, fruits, stems or the entire plants associated to a Natural Background category. Tree species like kaki or maple have generally more pictures and kind of views than herbaceous species like the mallow or the agrimony.

a nave way would therefore favor having such near-duplicate images in both the training and the test subsets, making the recognition much more easy. To avoid this bias, we therefore performed our random split at the observation level rather than at the image level thanks to associated metadata (observation id when available, author, date, etc.). The training data finally resulted in 20985 images while the test data resulted in 5092 images. According to similar concerns, the primary metric used to evaluate the submitted runs uses a two-stage average of raw image scores thanks to the users and observations ids associated to each test image. The raw image score itself is the inverse of the rank of the correct species in the list of retrieved species.

Participation and Results With 12 finalist groups over 9 countries and 33 submitted runs, the 2013 edition of the task confirmed its increasing attractiveness (respectively 10 and 11 groups crossed the finish line in 2011 and 2012) although its complexity was higher (with heterogeneous view types). Concerning the scan and scan-like images of leaves (called SheetAsBackground), the results of the 2013 task show that relatively high identification scores can be reach using leaf shape boundary features (between 0.6 and 0.5) but we cant notice a great step of improvement compared to the 2012 campaign. This can be explained by the fact that the queries were more difficult this year with more shadows, weaker lighting conditions, more old dried leaves and not so uniform background. Concerning the NaturalBackground category, results are as expected lower than the SheetAsBackground category. The highest scores reached equivalent values than the 2012 task, but without any human intervention in the workflow contrary to last year best runs involving some semi-automatic segmentation mechanisms. The detailed results by organ did show that most methods were clearly more accurate on the flower images rather than other organs. It corroborates a well-know usage of botanists for identifying plants and this is good news in a sense that computer vision methods go in the same direction. After flowers, there was no clear second best organ or view type. Bark images provided surprisingly good results relatively to the botanist knowhow on using bark morphology as an identification criterion. Identification results on the entire plant views are also rather surprising regarding their higher complexity and variability. Overall, an important remark is that the ranking of the runs did not change much from an organ to another one, fostering the idea that generic methods might solve heterogeneous fine-grained classification problems. Regarding metadata, one run did show that using the observation date complementary to the visual content was a simple and efficient way to obtain a gain of up to 5 points on the flower category (thanks to the relatively short flourishing season of many species). On the other side, the GPS information was not successfully exploited probably because the database doesnt contain dense enough observations to build an accurate geographic repartition of the species. With the emergence of more and more plant identification apps [9] [10], [11], [12] and the ecological urgency to build real-world and effective identification tools, we believe that the detailed results and conclusions of the task will be of high interest for the community [13].

2.3 The robot vision task

The Robot Vision task addresses two main problems related to semantic robot localization: place classification and object recognition. Participants are asked to answers the questions "where are you?" and "which object can you recognize in the scene?" when presented with a test sequence. Such test sequence contains depth and visual images acquired by a mobile robot with a RGB-D camera in a previously seen indoor environment.

Past Editions The Robot Vision task started in 2009 [14], with the main objective to compare different approaches to robot localization in a common scenario. The localization problem has always been managed from a semantic point of view, where no topological information is provided or required. Since its origin, new challenges have been introduced each new edition, from detection of unknown rooms [14], to generalization across floors [15, ?], to categorization problems [16] to multimodal data analysis [17]. At this fifth edition, 2013, the proposed challenge is the object recognition problem.

Objective and Task for the 2013 Edition For the 2013 edition, the semantic representation of the space is described by two elements which will determine the expected behaviour of people or robots in such scene. These two elements are: (1), the semantic category of the room (determines the activities we usually perform there, like Kitchen or Corridor) and (2), the list of objects the room contains (like Frigde or Desk). In a similar way topological localization (in conjunction with navigation and mapping) allows robots to move to a desired position, semantic localization is expected to provide robots with new capabilities. These capabilities are the identification of the most appropriate behaviour and the recognition of the objects that are suitable for interaction.
In this task edition, the relationship between room categories and objects is explicitly given. Using the labelling information, we can compute the conditional probability for a room category, given the list of objects in the scene $P(C = c_1|o_1, o_2, o_6)$ or vice versa $P(o_1|C = c1)$. This can be used to create a high level reasoning layer to be used in conjunction with low level classifiers. For example, the probability of detecting an Urinal in a Secretary is very low. Let us assume that we have classified a test frame as Secretary with high confidence but the object classifier cannot detect the presence or lack for the Urinal. In this case, the prior knowledge could be used to classify Urinal as not present. The use of this knowledge from participants is one of the goals of the challenge.
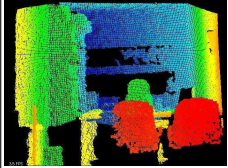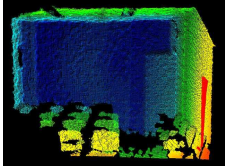
Task Description Participants are provided with two training sequences imaging all the rooms and object categories. They are expected to generate algorithms capable of providing information from test frames. Concretely, algorithms have to list all the objects that appear in the scene and classify the room category. The number of times a specific object appears in a frame is not relevant and, for each object, we have a binary problem. Room classification is a multi-class

problem.

Table 2 shows a global description for the task, where left columns correspond with the training stage of the challenge and the right one with the test. For both sequences, at each frame two different images are presented to the participants: a visual image and a point cloud. In the training stage, all the information for the room and the objects in the scene is provided. This information should be used by the participants teams to generate their algorithms. They have to classify the room category into one of the 10 available classes and say if each of the 8 possible objects are present or not. Due to wrong classifications will obtain negative scores, participants are allowed to not provide information about room category or object presence.

Table 2. Task Description.

Training		Test	
Visual and Depth Images		Visual and Depth Images	
Labels (provided)		Labels (required)	
Room Category	Objects	Room Category	Objects
Profesor Office	Extinguisher: NO	Class in Rooms	Extinguisher: Y/N/-?
	Computer: YES	or	Computer: Y/N/-?
	Printer: NO	Unknown	Printer: Y/N/-?
	Urinal: NO		Urinal: Y/N/-?
	Chair: YES		Chair: Y/N/-?
	Screen: NO		Screen: Y/N/-?
	Trash: YES		Trash: Y/N/-?
	Fridge: NO		Fridge: Y/N/-?

Performance Evaluation The proposals of the participants are compared using a score obtained from their submissions. The final score for a run will be the sum of all the scores obtained for the test frames included in the sequence. The following rules are used when calculating the final score for a frame:
Room Category (single multi-class problem)

– The room category has been correctly classified: +1.0 points
– The room category has been wrongly classified: -0.5 points
– The room category has not been classified: 0.0 points

Object (8 different binary problems)

– For each correctly classified object within the frame: +0.125 points

- For each misclassified object within the frame: -0.125 points
- For each object that was not classified: 0.0 points

The Data The dataset provided for the task consists of different sequences of depth (in Point Cloud Data (PCD) format [18]) and visual images acquired within a department building at the University of Alicante, Spain. Concretely, there are two labelled sequences for training, another labelled sequence provided for validation, and one unlabelled sequence for testing. Every image has been manually labelled with its corresponding room category and with a list of eight different objects to appear or not within it. The 10 different room categories are: corridor, hall, professorOffice, studentOffice, technicalRoom, toilet, secretary, visioconference, elevator area and warehouse. The 8 different objects are: extinguisher, computer, chair, printer, urinal, screen, trash and fridge. The frequency distribution for room categories and objects are depicted in Table 3 and Table 4 respectively.

Table 3. Frequency distribution of room categories for dataset sequences.

Room Category	Number of frames			
	Training 1	Training 2	Validation	Test
Corridor	891	1262	764	1317
Hall	103	228	000	297
ProfessorOffice	124	192	200	222
StudentOffice	155	276	282	318
TechnicalRoom	136	281	214	240
Toilet	121	242	188	198
Secretary	098	195	181	201
VisioConference	149	300	000	306
Warehouse	070	166	000	127
ElevatorArea	100	174	040	289
All	1947	3316	1869	3515

Corridor is the most common class in all sequences, due to the space distribution of the building used in the acquisition. This turns room classification into an unbalanced problem with higher probabilities for classifying frames as Corridor than for the rest of room categories. The validation sequence was released some months after the training sequences. The main objective of this sequence was to prevent the extreme lighting conditions of the test sequence. Due to it was acquired only in the first floor of the building, it does not contains any frame for three rooms: Warehouse, VisioConference and Hall.

Fig. 6 shows the same scene represented in three different sequences: training1 (top), validation (middle) and test (bottom). The scene was acquired using visual images (left) and point cloud data files (right). Training, validation and test sequences were acquired within the same building at two different floors but

Table 4. Frequency distribution of object presences or lacks for dataset sequences.

Room Category	Number of presences / lacks			
	Training 1	Training 2	Validation	Test
Extinguisher	259 / 1688	529 / 2787	286 / 1583	520 / 2995
Computer	289 / 1658	466 / 2850	416 / 1453	473 / 3042
Chair	470 / 1477	767 / 2549	567 / 1302	889 / 2626
Printer	210 / 1737	292 / 3024	255 / 1614	279 / 3236
Urinal	054 / 1893	110 / 3206	070 / 1799	090 / 3425
Screen	081 / 1866	190 / 3126	000 / 1869	151 / 3364
Trash	406 / 1541	451 / 2865	253 / 1616	662 / 2853
Fridge	057 / 1890	104 / 3212	099 / 1770	114 / 3401
All	1826 / 13750	2909 / 23610	1946 / 13006	3178 / 24942

with some variations in the lighting conditions (as can be observed in Fig. 6) and in the acquisition procedure (clockwise and counter clockwise, ground floor first or ground floor last). Participants were provided with running code for computing several feature descriptors [19–21] as well as SVM-based online [22, 23] and cure integration classifiers [24, 25].

Visual Image	Point Cloud File
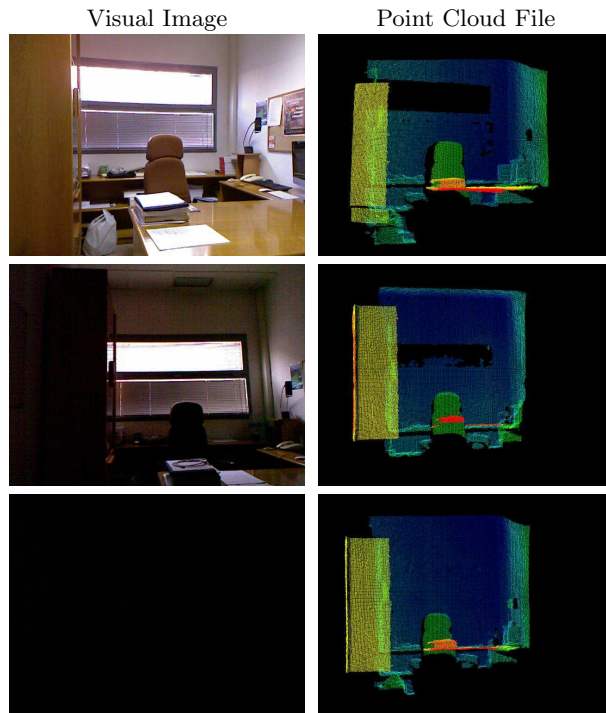
Fig. 6. Visual and 3D point cloud files for the same scene under different lighting conditions.

Participation and Results In 2013, 39 participants registered to the Robot Vision task but only 6 submitted, at least, one run accounting for a total of 16 different runs. These participants were:

- NUDT: National University of Defense Technology, Changsha, China.
- MIAR ICT: Beijing, China.
- MICA: Hanoi university of Science and Technology, Hanoi, Vietnam
- REGIM: University of Sfax National School of Engineers, Tunisia
- GRAM:University of Alcala de Henares, Spain
- SIMD: University of Castilla-La Mancha, Albacete, Spain.
  - Out of competition organizers contribution using proposed techniques

The scores obtained by all the submitted runs are shown in Table 5. The maximum score that could be achieved was 7030 and the winner (MIAR ICT) obtained a score of 6033.5 points. NUDT and SIMD teams ranked second and third respectively and their score was higher than 71% of the maximum score (the one obtained with the baseline system, SIMD result in the table).

Table 5. Overall ranking of the runs submitted by the participant groups to the 2013 Robot Vision task

Rank	Group Name	Score	% Max. Score
1	MIAR ICT	6033.500	85.83
2	MIAR ICT	5924.250	84.27
3	MIAR ICT	5924.250	84.27
4	MIAR ICT	5867.500	83.46
5	MIAR ICT	5867.000	83.46
6	NUDT	5722.500	81.40
7	SIMD*	5004.750	71.19
8	REGIM	4368.250	65.98
9	MICA	4479.875	63.73
10	REGIM	3763.750	53.54
11	MICA	3316.125	47.17
12	MICA	2680.625	38.13
13	GRAM	-487.000	<0.00
14	GRAM	-497.000	<0.00
15	GRAM	-497.000	<0.00
16	NUDT	-866.250	<0.00

* SIMD organizers submission was out-of-competition, it was provided to be considered a baseline score. The organizers only used the techniques proposed in the webpage of the Robot Vision challenge [2]. Concretely, PHOW [19] features were extracted from visual images and then, a Support Vector Machine was trained.

---

[2] http://www.imageclef.org/2013/robot

According to the obtained results we can conclude that the introduction of the object recognition task was not as challenging as we expected: most of the participants were able to identify those object properly. With respect to the scores obtained by the different runs, almost half of them improved the baseline results provided by the organizers, obtaining score higher than the 80% of the maximum score.

2.4 AMIA: the medical task

The main objective of the medical ImageCLEF task is to compare content-based image retrieval (CBIR) systems in medicine, and in particular to determine how associated cross-language text can be used in combination with CBIR to improve retrieval and ranking. ImageCLEFmed evaluates retrieval systems with visual, semantic and mixed topics in several languages using since 2008 a data collection from the biomedical literature.

Past Editions ImageCLEFmed started in 2004 with only an image-based retrieval task [26]. In 2005, an automatic annotation task was introduced [27]. The goal of this task was to find out how well the techniques can identify body orientation, body region, and biological system examined based on the images. The database consisted of 10,000 radiographs fully annotated with IRMA code, taken randomly from medical routine. Between 2006 and 2009, ImageCLEFmed kept these two tasks in similar formats format but using larger and more complex databases each year [28–32]. From 2008 to 2010, the database contained images from articles published in Radiology and Radiographics including the text of the captions and a link to the html of the full text articles. In 2009, a lung nodule detection task was tested. The goal of this task was to compare the performance of lung nodule detection techniques with a gold standard of manually identified nodules. The data for this task was a subset of the LIDC (Lung Image Database Consortium) database. From 2010 to 2012, there were three types of task: the traditional image-based retrieval, modality classification and case-based retrieval [33–35]. The modality classification task was introduced since previous studies have shown that imaging modality is an important aspect of the image for medical retrieval. Using the modality classification the search results can be improved significantly. In 2010, the images had to be classified into one of 8 modalities (CT, MR, XR, etc.); in 2011 into 18 and in 2012-2013 into 31. In the case-based retrieval task, a case description, with patient anamnesis, limited symptoms and test results including imaging studies is provided (but not the final diagnosis). The goal is to retrieve cases including images that are useful for a differential diagnosis or even match the exact diagnosis of the query.

Objective and Task for 2013 Edition In 2013, the 10th year of the medical task is celebrated [36]. The ImageCLEFmed meeting will be organized at the annual AMIA meeting in the form of a workshop. This means that the workshop will be organized outside of Europe for the first time. ImageCLEFmed

is running in a similar format as in 2012 but with a new task, the compound figure separation that became important as a large fraction of around 40% of the database of PubMed Central used contain compound figures and the sub images are otherwise not accessible for research. Another novelty in 2013 is that the modality classification task includes a large amount of compound images to make the task more difficult and realistic. The following tasks were offered in 2013:

- Modality Classification: In user-studies, 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. However, this modality is typically extracted from the caption and is often not correct or present. Studies have shown that the modality can be extracted from the image itself using visual features. Additionally, using the modality classification, the search results can be improved significantly. In 2013, a larger number of compound figures will be present making the task significantly harder but corresponding much more to the reality of biomedical journals.
- Compound figure separation: As up to 40% of the figures in PubMed Central are compound figures, a major step in making the content of the compound figures accessible is the detection of compound figures and then their separation into sub figures that can subsequently be classified into modalities and made available for research. The task makes available training data with separation labels of the figures, and then a test data set where the labels were made available after the submission of the results.
- Ad-hoc image-based retrieval: This is the classic medical retrieval task, similar to those in organized since 2004. Participants were given a set of 30 textual queries with 2-3 sample images for each query. The queries were classified into textual, mixed and semantic, based on the methods that are expected to yield the best results.
- Case-based retrieval: This task was first introduced in 2009. Unlike the ad-hoc task, the unit of retrieval here is a case, not an image. For the purposes of this task, a "case" is a PubMed ID corresponding to the journal article. In the results submissions the article DOI should be used as several articles do neither have PubMed IDs nor Article URLs.

The medical image classication and retrieval tasks in 2013 cover image modality classication, compound image separation and image retrieval with visual, semantic and mixed topics in several languages using a data collection from the biomedical literature.

Participation and Results In total over 60 groups registered for the medical tasks and obtained access to the data sets.10 of the registered groups submitted results to the medical tasks with a total of 166 valid runs submitted. 8 groups participated in the modality classification task with 51 runs; 3 groups participated in the compound figure separation task with 4 runs; 9 groups participated

in the image retrieval task with 66 runs and 7 groups participated in the case-based retrieval task with 45 runs. As in previous years, the largest number of runs was submitted for the image-based retrieval task although the number submitted runs at the modality classication task increased to 51 (43 in 2012 and 34 in 2011). There are still different situations as to whether visual, textual or combined techniques perform better depending on the task. For further information you can see the ImageCLEFmed overview [36].

3   Conclusion

This paper presented an overview of the activities in the 2013 edition of the ImageCLEF lab. The sustained interest in the lab, witnessed by the growing number of registration and the sustained number of groups actually participating to the lab, make ImageCLEF an important resource in the multi lingual image annotation and retrieval research landscape. The ever growing amount of data available through the internet, and the growing demand of tools for accessing and exploiting them, will become one of the key focus for the 2014 edition of ImageCLEF, where we look forward to welcome back the medical task under the ImageCLEF umbrella.

3.1   Acknowledgments

References

1. Muller, H., Clough, P., Deselaers, T., Caputo, B.: ImageCLEF: experimental evaluation in visual information retrieval. Springer (2010)
2. Tsikrika, T., de Herrera, A.S., Mller, H.: Assessing the scholarly impact of imageclef. In: Cross Language Evaluation Forum (CLEF 2011). Lecture Notes in Computer Science (LNCS), Springer (2011)
3. Huiskes, M., Lew, M.: The MIR Flickr retrieval evaluation. In: Proceedings of the 10th ACM Conference on Multimedia Information Retrieval, Vancouver, BC, Canada (2008) 39–43
4. Huiskes, M., Thomee, B., Lew, M.: New trends and ideas in visual concept detection. In: Proceedings of the 11th ACM Conference on Multimedia Information Retrieval, Philadelphia, PA, USA (2010) 527–536
5. Villegas, M., Paredes, R.: Overview of the ImageCLEF 2012 Scalable Web Image Annotation Task. In: CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy (2012)
6. Zellhöfer, D.: Overview of the Personal Photo Retrieval Pilot Task at ImageCLEF 2012. In: CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy (2012)

7. Villegas, M., Paredes, R., Thomee, B.: Overview of the ImageCLEF 2013 Scalable Concept Image Annotation Subtask. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes, Valencia, Spain (2013)
8. Zellhöfer, D.: Overview of the ImageCLEF 2013 Personal Photo Retrieval Subtask. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes, Valencia, Spain (2013)
9. : Leafsnap (2011)
10. : Plantnet (2013)
11. : Mobile flora (2013)
12. : Folia (2012)
13. Goëau, H., Bonnet, P., Joly, A., Bakic, V., Boujemaa, N., Barthelemy, D., Molino, J.F.: The imageclef 2013 plant identification task. In: ImageCLEF 2013 Working Notes. (2013)
14. Pronobis, A., Xing, L., Caputo, B.: Overview of the clef 2009 robot vision track. Springer (2010) 110–119
15. Pronobis, A., Caputo, B.: The robot vision task. In Muller, H., Clough, P., Deselaers, T., Caputo, B., eds.: ImageCLEF. Volume 32 of The Information Retrieval Series. Springer Berlin Heidelberg (2010) 185–198
16. Pronobis, A., Christensen, H., Caputo, B.: Overview of the imageclef@ icpr 2010 robot vision track. Recognizing Patterns in Signals, Speech, Images and Videos (2010) 171–179
17. Martinez-Gomez, J., Garcia-Varea, I., Caputo, B.: Overview of the imageclef 2012 robot vision task. In: CLEF 2012 working notes. (2012)
18. Rusu, R., Cousins, S.: 3d is here: Point cloud library (pcl). In: Robotics and Automation (ICRA), 2011 IEEE International Conference on, IEEE (2011) 1–4
19. Bosch, A., Zisserman, A., Munoz, X.: Image classification using random forests and ferns. In: International Conference on Computer Vision, Citeseer (2007) 1–8
20. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 1., Ieee (2005) 886–893
21. Linde, O., Lindeberg, T.: Object recognition using composed receptive field histograms of higher dimensionality. In: Proc. ICPR, Citeseer (2004)
22. Orabona, F., Castellini, C., Caputo, B., Luo, J., Sandini, G.: Indoor place recognition using online independent support vector machines. In: Proc. BMVC. Volume 7. (2007)
23. Orabona, F., Castellini, C., Caputo, B., Jie, L., Sandini, G.: On-line independent support vector machines. Pattern Recognition **43** (2010) 1402–1412
24. Orabona, F., Jie, L., , Caputo, B.: Online-Batch Strongly Convex Multi Kernel Learning. In: Proc. of Computer Vision and Pattern Recognition, CVPR. (2010)
25. Orabona, F., Jie, L., Caputo, B.: Multi kernel learning with online-batch optimization. Journal of Machine Learning Research **13** (2012) 165–191
26. Clough, P., Müller, H., Sanderson, M.: The CLEF 2004 cross–language image retrieval track. 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
27. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., Hersh, W.: The CLEF 2005 cross–language image retrieval track. In: Cross Language Evaluation Forum (CLEF 2005). Lecture Notes in Computer Science (LNCS), Springer (2006) 535–557

28. Müller, H., Deselaers, T., Lehmann, T., Clough, P., Kim, E., Hersh, W.: Overview of the ImageCLEFmed 2006 medical retrieval and medical annotation tasks. In: CLEF 2006 Proceedings. Volume 4730 of Lecture Notes in Computer Science (LNCS)., Alicante, Spain, Springer (2007) 595–608

29. 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

30. Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Radhouani, S., Bakke, B., Kahn, J.C.E., Hersh, W.: Overview of the clef 2009 medical image retrieval track. In: Proceedings of the 10th international conference on Cross-language evaluation forum: multimedia experiments. CLEF'09, Berlin, Heidelberg, Springer–Verlag (2010) 72–84

31. Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Radhouani, S., Bakke, B., Kahn, J.C.E., Hersh, W.: Overview of the clef 2009 medical image retrieval track. In: Proceedings of the 10th international conference on Cross-language evaluation forum: multimedia experiments. CLEF'09, Berlin, Heidelberg, Springer–Verlag (2010) 72–84

32. Tommasi, T., Caputo, B., Welter, P., Guld, M., Deserno, T.: Overview of the clef 2009 medical image annotation track. In: Proceedings of the 10th international conference on Cross-language evaluation forum: multimedia experiments. CLEF'09, Berlin, Heidelberg, Springer–Verlag (2010) 85–93

33. 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)

34. 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). (2011)

35. Müller, H., García Seco de Herrera, A., Kalpathy-Cramer, J., Demner Fushman, D., Antani, S., Eggel, I.: Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In: Working Notes of CLEF 2012 (Cross Language Evaluation Forum). (2012)

36. García Seco de Herrera, A., Kalpathy-Cramer, J., Demner Fushman, D., Antani, S., Müller, H.: Overview of the ImageCLEF 2013 medical tasks. In: Working Notes of CLEF 2013 (Cross Language Evaluation Forum). (2013)