ImageCLEF 2014: Overview and Analysis of the Results

Barbara Caputo[1], Henning Müller[2], Jesus Martinez-Gomez[3], Mauricio Villegas[4], Burak Acar[5], Novi Patricia[6], Neda Marvasti[5], Suzan Üsküdarlı[5], Roberto Paredes[4], Miguel Cazorla[7], Ismael Garcia-Varea[3], and Vicente Morell[7]

[1] University of Rome La Sapienza, Italy
[2] University of Applied Sciences Western Switzerland in Sierre, Switzerland
[3] University of Castilla-La Mancha, Spain
[4] PRHLT, Universitat Politècnica de València, Spain
[5] Bogazici University, Turkey
[6] Idiap Research Institute, Switzerland
[7] University of Alicante, Spain

Abstract. This paper presents an overview of the ImageCLEF 2014 evaluation lab. Since its first edition in 2003, ImageCLEF has become one of the key initiatives promoting the benchmark evaluation of algorithms for the annotation and retrieval of images in various domains, such as public and personal images, to data acquired by mobile robot platforms and medical archives. Over the years, by providing new data collections and challenging tasks to the community of interest, the ImageCLEF lab has achieved an unique position in the image annotation and retrieval research landscape. The 2014 edition consists of four tasks: domain adaptation, scalable concept image annotation, liver CT image annotation and robot vision. This paper describes the tasks and the 2014 competition, giving a unifying perspective of the present activities of the lab while discussing future challenges and opportunities.

1 Introduction

Since its first edition in 2003, the ImageCLEF lab has aimed at providing an evaluation forum for the language independent annotation and retrieval of images [19]. Motivated by the need to support multilingual users from a global community accessing the ever growing body of visual information, 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 systems operating in monolingual, language-independent and multi-modal contexts, providing reusable resources for benchmarking. To meet its objectives, ImageCLEF organises tasks that benchmark the annotation and retrieval of diverse images such as general photographic, medical images and adapting knowledge across different domains, as well as domain-specific tasks such as robot vision. These tasks aim to support and promote research that addresses key challenges in the field. 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. 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 [32].

The remainder of this paper is organized as follows: section 2 describes the four subtasks of the 2014 edition: the domain adaptation task (section 2.1), the scalable concept image annotation task (section 2.2), the liver CT image annotation task (section 2.3) and the robot vision task (section 2.4). We conclude with an overall discussion, and pointing towards the challenges ahead and possible new directions for ImageCLEF 2015.

2 ImageCLEF 2014: the tasks, the data and participation

The 2014 edition of ImageCLEF consisted of four main tasks: the domain adaptation task, the scalable concept image annotation task, the liver CT image annotation task and the robot vision 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 liver CT annotation 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 acquirednand thus noisy labelled samples,
- scientific multimedia data management through the particular case of liver CT image annotation,
- the ability of generic annotation algorithms to adapt robustly and effectively across domains, 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 and/or within the state of the art in each respective field, of its 2014 objective and task, and of the task participation and relative results.

2.1 Domain Adaptation Task

The amount of freely available and annotated image collections has dramatically increased over the last years, thanks to the diffusion of high-quality cameras and also to the introduction of new and cheap annotation tools such as Mechanical Turk. Attempts to leverage over and across such large data sources has proved

challenging. Indeed, tools like Google Goggle are able to relaibly recognize limited classes of objects like books or wine labels, but are not able to generalize across generic objects like food items, clothing items and so on. Several authors showed that, for a given task, training on a dataset (e.g. Pascal VOC 07) and testing on another (e.g. ImageNet) produces very poor results, although the set of depicted object categories is the same [26,31]. In other words, existing object categorization methods do not generalize well across databases.

This problem is known in the literature as the domain adaptation challenge. Addressing this issue would have a tremendous impact on the generality and adaptability of any vision-based annotation system. Current research in domain adaptation focuses on a scenario where

- (a) the prior domain (source) consists of one or a maximum of two databases;
- (b) the labels between the source and the target domain are the same, and
- (c) the number of annotated training data for the target domain are limited.

The goal of the Domain Adaptation Task, initiated in 2014 under the ImageCLEF umbrella [2], is to push the state of the art in domain adaptation towards more realistic settings, relaxing these assumptions. Our ambition is to provide, over the years, stimulating problems and challenging data collections that might stimulate and support novel research in the field.

Objective and Task for the 2014 Edition In the 2014 version (first edition) of the Domain Adaptation Task, we focused on the number of sources available to the system. Current experimental settings, widely used in the community, consider typically one source and one target [26], or at most two sources and one target [9,30]. This scenario is unrealistic: with the wide abundance of annotated resources and data collections that are made available to users, and with the fast progress that is being made in the image annotation community, it is likely that systems will be able to access more and more databases and therefore to leverage over a much larger number of sources than two, as considered in the most challenging settings today.

To push research towards more realistic scenarios, the 2014 edition of the domain adaptation task has proposed an experimental setup with four sources, where such sources were built by exploiting existing available resources. Participants were thus requested to build recognition systems for the target classes by leveraging over such source knowledge. We considered a semi-supervised setting, i.e. a setting where the target data, for each class, is limited but annotated.

Specifically, to define the source and target data, we considered five publicly available databases:

- the *Caltech-256* database, consisting of 256 object categories, with a total of 30.607 images;
- the *ImageNet ILSVRC2012* database, organized according to the WordNet hierarchy, with an average of 500 images per node;
- the *PASCAL VOC2012* database, an image data set for object class recognition with 20 object classes;

– the *Bing* database, containing all 256 categories from the Caltech-256 one, and augmented with 300 web images per category that were collected through textual search using Bing;
– and the *SUN* database, a scene understanding database that contains 899 categories and 130.519 images.

We then selected twelve classes, common to all the datasets listed above: aeroplane, bike, bird, boat, bottle, bus, car, dog, horse, monitor, motorbike, and people. Figure 1 illustrates the images contained for each class in each of the considered datasets. As sources, we considered 50 images represented the classes listed above from the databases Caltech-256, ImageNet, PASCAL and Bing. The 50 images were randomly selected from all those contained in each of the data collection, for a total of 600 images for each source. As target, we used images taken from the SUN database for each class. We randomly selected 5 images per class for training, and 50 images per class for testing. These data were given to all participants as validation set. The test set consisted of 50 images for each class, for a total of 600, manually collected by us using the class names as textual queries with standard search engines.

Instead of making the images directly available to participants, we decided to release pre-computed features only, in order to keep the focus on the learning aspects of the algorithms in this year's competition. Thus, we represented every image with dense SIFT descriptors (PHOW features) at points on a regular grid with spacing 128 pixels [1]. At each grid point the descriptors were computed over four patches with different radii, hence each point was represented by four SIFT descriptors. The dense features have been vector quantized into 256 visual words using k-means clustering on a randomly chosen subset of the Caltech-256 database. Finally, all images were converted to 2×2 spatial histograms over the 256 visual words, resulted in 1024 feature dimension. The software used for computing such features is available at www.vlfeat.org.

Participation and Results While 19 groups registered to the domain adaptation task to receive access to the training and validation data, only 3 groups eventually submitted runs: the XRCE group, the Hubert Curien Lab group and the Idiap group (organizers). They submitted the following algorithms:

– the XRCE group submitted a set of methods based on several heterogeneous methods for domain adaptation, of which predictions were subsequently fused. By combining the output of instance based approaches and metric learning one with a brute force SVM prediction, they obtained a set of heterogeneous classifiers all producing class prediction for the target domain instances. These were combined through different versions of majority voting in order to improve the overall accuracy.
– The Hubert Curien Lab group did not submit any working notes, neither sent any detail about their algorithm. We are therefore not able to describe it.

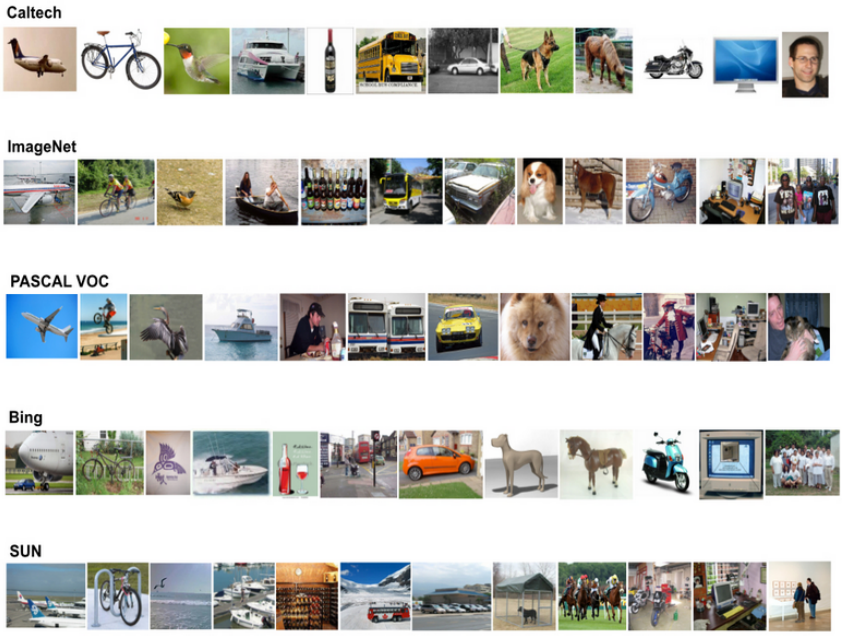


Fig. 1: Exemplar images for the 12 classes from the five selected public databases.

– The Idiap group submitted a baseline run using a recently introduced learning to learn algorithm [21]. The approach considers source classifiers as experts, and it combines their confidence output with a high-level cue integration scheme, as opposed to a mid-level one as proposed in [10]. The algorithm is called High-level Learning to Learn (H-L2L). As our goal was not to obtain the best possible performance but rather to provide an off the shelf baseline against which to compare results of the other participants, we did not perform any parameter tuning.

Table 1 reports the final ranking among groups. We see that XRCE obtained the best score, followed by the Hubert Curien lab. The Idiap baseline obtained the worst score, clearly pointing towards the importance of parameter selection in these kind of benchmark evaluations.

For the complete results, details and analysis, please refer to the task overview paper [3].

2.2 Scalable Concept Image Annotation task

Automatic concept detection within images is a challenging research problem, and as of today unsolved. Despite considerable research efforts the so-called semantic gap has not been successfully breached, in terms of being able to detect

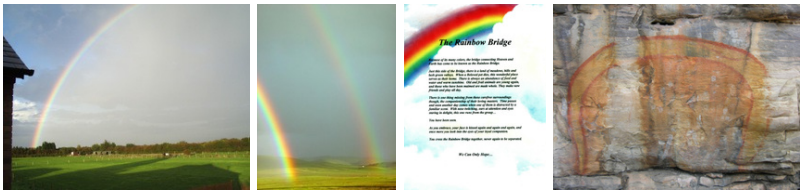
Rank	Group	Score
1	XRCE	228
2	Hubert Curien Lab Group	158
3	Idiap	45

Table 1: Ranking and best score obtained by the three groups that submitted runs.

semantic concepts within any kind of imagery for any kind of concept as accurately as humans can. Furthermore, the greatest achievements in this research area are characterized by the reliance on clean hand labeled training data, a fact that greatly limits the scalability of the developed approaches. ImageCLEF's Scalable Concept Image Annotation task aims to advance the state of the art in image concept detection by acting as a platform to foster interaction and collaboration between researchers and by providing a realistic and challenging benchmark with a particular incentive for the development of technologies that are able to scale concept-wise without the requirement of large amounts of human effort.

Past Editions The Scalable Concept Image Annotation task is a continuation of the general image annotation and retrieval task that 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. In its current form, the 2014 Scalable Concept Image Annotation task is its third edition, having been organized in 2012 [35] and 2013 [37] as subtasks of the the general image annotation and retrieval task. This is the first year in which this scalable annotation aimed benchmark has been organized as a standalone main task.

Objective and Task for the 2014 Edition 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 grow. However, images for any topic can be cheaply gathered from the 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. Figure 2 shows examples of typical images found by querying search engines. As can be seen, the data obtained are useful and furthermore a wider variety of images is expected, not only photographs, but also drawings and computer generated graphics. Likewise there are other resources available that can help to determine the relationships between text and semantic concepts, such as dictionaries or ontologies.



(a) Images from a search query of "rainbow".

(b) Images from a search query of "sun".

Fig. 2: Example of images retrieved by a commercial image search engine.

The goal of this task 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, other similar image datasets and any other automatically obtained resources. There were several differences in this task with respect to the previous edition. First the list of concepts to detect was increased from 116 to 207, but most importantly the concepts in the test set not seen during development increased from 21 to 100. Another difference was that each image of the test set had its own list of concepts to detect, so not all images had to be annotated for the 207 concepts. This permitted among other things to have exactly the same 2013 test set as a subset, and also to have subsets of images in which all of the concepts to detect were not seen during development. A final difference to mention was that the amount of training data provided was doubled.

The data used in this task was similar to the one from last year [37], in fact half of the training data provided were exactly the same. The training set was composed of 500,000 samples each of which included: the raw image, seven types of precomputed visual features and four types of textual features. These training images were obtained from the web by querying popular image search engines. The development and test sets had 1,940 and 7,291 samples, respectively, which only included the visual features and the corresponding hand labeled concepts ground truth. The ground truth for the test set was not released, it was kept secret so that the participants had to submit the annotation results which were then analyzed by the task organizers. For further details, please refer to the task overview paper [36].

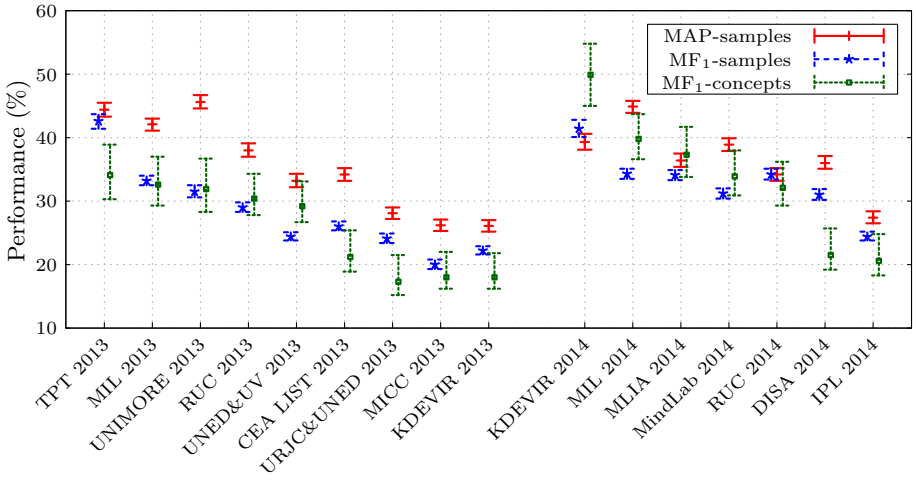


Fig. 3: The three performance measures for the best submission of each group for both this and last year's edition of the task. The results for both years are for the same test set (since the 2013 test set was included this year as a subset). Error bars correspond to the 95% confidence intervals computed using Wilson's method.

Participation and Results Generally speaking, the participation was excellent, although there was a slight decrease in participation with respect to last year. In total, 11 groups took part in the task and submitted overall 58 system runs. Among the 11 participating groups, only 7 submitted a working notes paper, thus only for these there were specific details of their systems available. Last year the participation was 13 groups, 58 runs and 9 papers.

Last year it was decided that the ground truth for the test set would not be released, so that the same data could be reused every year and be able to observe the evolution of the developed systems overtime. Figure 3 presents the results, both for this and last year's edition of the task, in each case for the best system of each group that submitted a paper. The graph includes the three performance measures that were used to judge the systems, which were: the Mean Average Precision (MAP) computed for the samples, and the mean F_1-measure computed both for the samples and for the concepts. In the figure it can be observed that in general this year the participants obtained better systems, most of them achieving performances over 30% for the three measures.

One shortcoming found last year was that the number of concepts in the test set not seen during development was too small, so when comparing the performance of the different systems the confidence intervals were too wide, making it difficult to derive adequate conclusions. For this reason, an objective for this year was to increase the number of unseen concepts, and thus these were

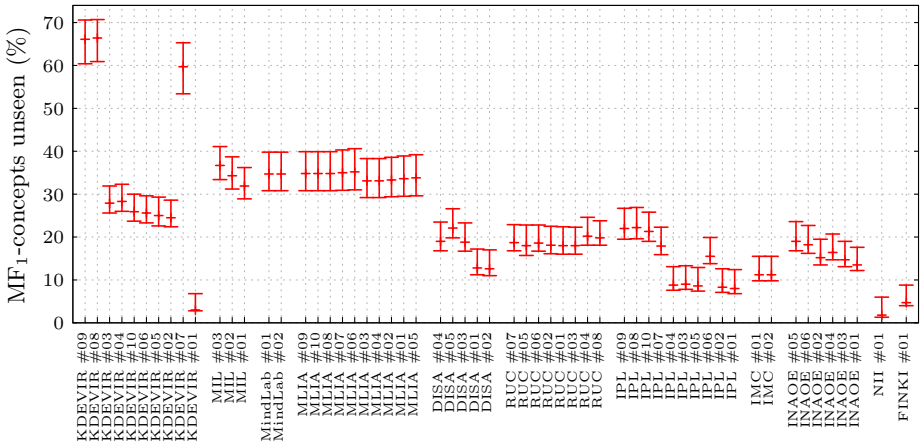


Fig. 4: The results for the test set for all of the submissions using the MF_1-concepts performance measure although only considering the concepts that were not seen during development. Error bars correspond to the 95% confidence intervals computed using Wilson's method.

the most interesting results obtained this year, which are presented in Figure 4. In these results it can be clearly observed which systems outperform the others.

Last year the best team TPT [27] obtained very good results by using learning techniques that take into account context, effectively finding a way to exploit the information available in the noisy webpage data. This year the best team KDEVIR [25] decided to follow the same line as TPT, however, on top of that they have developed techniques for automatically building ontologies for the concepts and using these both in training phase for better selecting the images used for optimizing the classifiers and in the testing phase for taking into account the relationships between the concepts.

It is curious to observe that the performance of the best system is significantly better for the unseen concepts ($\approx 65\%$) than overall ($\approx 50\%$). This is possibly because the unseen concepts are relatively easier, however when comparing with other systems it can indicate the importance of using automatically generated ontologies in this challenge.

For the complete results, details and analysis, please refer to the task overview paper [36].

2.3 Liver CT Image Annotation Task

Medical images present unique challenges in comparison to other images. A significant part of the medical image analysis tasks deal with a set of rather similar images, such as abdominal CT images, where the analysis is based on subtle differences between images. In a conventional setting, these subtle differences,

such as the texture observed in the parenchyma of a liver, are observed by experts and translated into the medical vocabulary using medical terminology that constitutes image anotation. An annotation facilitates the high-level processing and communication of medical evidence derived from the images. In recognition of its importance, several standard terminologies are being developed/used, such as SNOMED-CT[8] (Systemized Nomenclature in Medicine), RadLex[9] (Radiology Lexicon), NCBO[10], UMLS[11] (Unified Medical Language System), etc. Despite its advantages, an expert annotation is a labor intensive task that can be performed by qualified individuals only and must be consistent among different individuals, sites, countries, etc. Hence, a key challenge in expert annotation is to translate computer generated objective low-level image observations (CoG) to high level semantic descriptions (ie. annotations) that comply with a standard terminology of choice. Such an automatic medical image annotation system can facilitate effective multi-site communication of medical information, semantic search and retrieval in (multi-site) medical databases, human-interpretable computer aided diagnosis, computer aided reporting, etc. The "Liver CT Image Annotation Task", introduced for the first time in ImageCLEF 2014, focused on the aforementioned challenge and is restricted to the liver CT image annotation, as a pilot application domain.

Previous Work The automatic image annotation methods in the literature can be categorized as the classification based approaches [38] and the Bayesian methods [34]. In the classification based approaches, the annotation problems is addressed as a multi-class classification problem. Here, every semantic concept is treated as a class and a set of binary classification models are used to give yes-or-no votes. Conventional classifiers for this task include Support Vector Machines (SVM), Artificial Neural Networks (ANN), Decision Trees (DT), and Random Forest(RF). The majority of the proposed systems fall in this category. Shi et al.[29] trained an SVM with radial basis function kernel to annotate a set of images. They trained the SVM classifiers using the image features for every concept. Each classifier generates a probability value, fused with other SVM outputs producing the final label of that feature by applying majority voting. Goh et al. [8] pursued a similar approach and used a 3-level model to annotate a set of images. They used different sets of classifiers, estimated the decision for each set using majority voting and finally fused all decisions to get the final label. Qj et al. [24] have also used a three level classifier for two sets of SVM classifiers. The first group uses global features and the second group employs local features. Mueen et al. [18] have implemented the annotation using three-hierarchy-level SVM classification on X-ray images. Devrim et al. [33] used two approaches to automatically annotate X-ray images. In the first approach, they used a single SVM with 1-vs-all multi-class model and Gaussian radial basis function. In the

[8] http://www.ihtsdo.org/snomed-ct/
[9] http://www.radlex.org
[10] http://www.bioontology.org
[11] http://www.nlm.nih.gov/research/umls/

second approach, they used separate SVM classifiers for each label and finally fused their classification results. Frat et al. [5] , Kim et al. [11], and Park et al. [20] have employed ANN to perform automatic image annotation. Data training with ANN algorithms is time-consuming, but they can learn multiple classes simultaneuously. The number of layers and nodes of ANNs influence the performance. DT was employed by Friedl M.A. et al. [7], Wong R.C.F et al. [39], and Sethi I.K. et al. [28] to annotate land covers in the remote sensed data, real-world web images, and outdoor images into semantic concepts, respectively. A DT is a multi-stage decision making / classification tool, in which we have a set of root nodes, a set of terminal nodes, and a set of leaf nodes. DT divides the data into smaller non-overlapping subsets, according to if-then-rules. Byoung et al. [12] used a combination of RFs and wavelet-based center symmetric-local binary patterns for medical image classification to perform multiple keyword annotations. It has been shown that classification using RFs is much faster than SVMs. Bayesian probability rules can also be used to classify and annotate images. Particularly, in the training step the conditional probability of an image, being labeled by every class, is calculated using some parametric [40],[6] or non-parametric [34] methods. Then in test phase, the class/label of the image is defined by maximizing the posterior (MAP) criterion.

Objective and Task for the 2014 Edition The participants were given a training set of 50 cropped liver CT images together with the liver, vessel and lesion masks, a set of 60 computer generated features (CoG) and a set of 73 manual semantic annotations (UsE) regarding the liver, the vessels and one selected lesion. The UsE features were generated by an expert radiologist as part of the CaReRa[12] (Case Retrieval in Radiology) project, using the opensource ONLIRA (Ontology Of Liver For Radiology) [13]. The test set had 10 cases, with all types of data available in the training set except the UsE features. The particpants were asked to estimate the missing 73 UsE features. They were allowed to use any subset or superset of the provided CoG features, giving them the option to compute and use any additional low-level features that they may extract from the CT images and the masks. The evaluation was based on the *Completeness* (defined as the percentage of all 73 UsE features that were estimated) and *Accuracy* (defined as the percentage of the estimated UsE features that were correct), geometric mean of which was used as the *Total Score*. Ideally, all metrics would be 1.00.

Participation and Results Three groups participated in this task: BMET (University of Sydney), CASMIP (The Hebrew University of Jerusalem), piLab-VAVlab (Bogaziçi University).

Table 2 lists the results of all runs submitted. compares the results of different runs in predicting different groups of UsE features. We divide UsE features

[12] TUBITAK-ARDEB grant no 110E264, PI: Burak Acar, PhD.
http://www.vavlab.ee.boun.edu.tr/pages.php?p=research/CARERA/carera.html

into 5 groups: liver, vessels and three lesion groups with area, lesion and component concepts. Results show that most of the methods predicted the vessel UsE features completely.

Group name	Run	Completeness	Accuracy	Total Score	method used	feature used
BMET	run1	0.98	0.89	0.935	SVM-linear	CoG
BMET	run2	0.98	0.90	0.939	SVM-linear	CoG+
BMET	run3	0.98	0.89	0.933	SVM-RBF	CoG
BMET	run4	0.98	0.90	0.939	SVM-RBF	CoG+
BMET	run5	0.98	0.91	0.947	IR-noFS	CoG
BMET	run6	0.98	0.87	0.927	IR-noFS	CoG+
BMET	run7	0.98	0.91	0.947	IR-FS	CoG
BMET	run8	0.98	0.87	0.926	IR-FS	CoG+
CASMIP	run1	0.95	0.91	0.93	LDA+KNN	CoG+
piLabVAVlab	run1	0.51	0.39	0.45	TF-KL	CoG
piLabVAVlab	run2	0.51	0.89	0.677	TF-EUC	CoG
piLabVAVlab	run3	0.51	0.88	0.676	TF-KL	CoG

Table 2: Results of the runs of Liver CT annotation task. CoG: The provided CoG features. CoG+: The extended set of CoG feature (Each group, when applicable, used a their own extension which is explained in the text). SVM: Support Vector Machine. IR-(no)FS: Image Retrieval w/o Feature Selection. TF: Tensor Factorization

The BMET group pursued two approaches: The classifier based approach and the retrieval based approach. They repeated all experiments with the provided CoG features and with an extended set of features where they added the bag of visual words (BoWD) to the CoG set. The classification based approach utilized a bank of SVMs, one for each UsE feature to be predicted, in two stages. The first stage used 1-vs-all classifiers, where, for a given concept, each label is learned against all other possible labels of that concept. In case the first stage cannot identify a single label, the second stage is applied where 1-vs-1 classifiers are used to break the tie. Linear and RBF kernels are used in the SVM classifiers. The retrieval based approach aimed at identifying n ($n = 10$) taining cases that are closest to the test case in terms of the Euclidean distance in the feature space. A weighted voting scheme is applied to determine the UsE features of the test case using those of the identified training cases. The retrieval is applied with and without feature selection. Extending the CoG feature set did not improve the results significantly, it even decreased accuracy in the retrieval based approach. On the other hand, the retrieval based approach performed better than the classification based one.

The CASMIP group pursued a classification based approach in a lower dimensional feature space. They excluded 21 CoG features but added 9 new features describing the gray level of liver, lesion and the lesion boundary. They used 4 different classifiers in their experiments: Linear Discriminant Analysis

(LDA), Logistic Regression (LR), K-Nearest Neighbours (KNN) and Support Vector Machines (SVM). All classifiers are trained on individual UsE features to be predicted and the best performing one of the 4 classifiers, is chosen for each UsE features separately. It turned out that LDA and KNN were the best performing classifiers for the majority of UsE features. Cluster_Size, Lesion_Lobe and Lesion_Segment were deterministically determined from the CoG features.

The piLabVAVlab group pursued a drastically different approach and assessed the use of Generalized Coupled Tensor Factorization (GCTF) [41]. The GCTF is a general framework where a high order tensor representation of the conditional probabilities (between CoG and UsE features) are used. KL-divergence and Euclidean distance is used in the tensor factorization problem modeling. Though the input UsE features are categorical, the output of GCTF is real valued, hence the method requires approapriate thresholding, which affects the results (as seen among different runs of this group). Furthermore, the piLabVAVlab group attempted to predict only the UsE features with 4 or 2 labels (categories), which set their completeness upper limit to 0.51. Their accuracy was not significantly different than the other groups', suggesting the GCTF as a promising method, which is totally blind to the domain knowledge.

Despite the small dataset size, the "Liver CT Image Annotation Task", introduced this year, demonstrated the feasibility of automatic medical image annotation from low level image features by means of retrieval, supervised machine learning and GCTF. None of the methods, specifically the GCTF, utilized the domain knowledge as represented with an ontology. It can be conjectured that using the domain knowledge would improve the results even further, paving the way for automatic radiology reporting and semantic search using low-level image features.

2.4 Robot Vision Task

The Robot Vision task addresses two problems in parallel: room classification and object recognition. Participants of the challenge are asked to classify rooms on the basis of visual and depth images captured by a Kinect sensor mounted on a mobile robot. Moreover, participants are also asked to detect the appearance or lack of a list of previously defined objects.

Past Editions The first edition of the Robot Vision task started in 2009 [22], and since its origin, it has addressed the problem of place classification with application to robotics. This problem consists in answering the question "where am I?" from a semantic point of view. That is, using semantic information like I am in the office instead of metric one.

The procedure of the task has maintained similar from the first edition. Firstly, the organizers define the problem, the performance evaluation procedure, and release images annotated with semantic information for training. Participants are then expected to start developing their proposals using the provided information. Some time later, an annotated validation sequence is released. This

sequence allows participants to estimate whether their algorithms perform well when facing new images not previously seen. Finally, an unannotated test sequence is released and participants have some days to process it. As a result of this processing, a submission file with the obtained annotations has to be uploaded. All the participant submissions are then evaluated (using the previously presented procedure) and ranked to determine the winner of the task.

Each new edition of the Robot Vision task has introduced new changes in the data provided to the participants as well as for the requested information. Some of the most important variations are enumerated in the following: the use of stereo images (2010@ICPR [23]), the inclusion of depth information (2012@ImageCLEF [15]), point cloud representation for depth information and object recognition problem (2013@ImageCLEF [16]).

Objectives and Task for the 2014 Edition The sixth edition of the Robot Vision challenge [17] focuses on the use of multimodal information (visual images and point cloud files) with application to semantic localization and object recognition. It addresses the problem of robot localization in parallel to object recognition from a semantic point of view, and with a special interest in the capability of generalization. Both problems are inherently related: the objects present in a scene can help determine the room category and vice versa.

Participants were provided with visual and depth images in Point Cloud Data (PCD) format. In addition to all the image sequences, we created a Matlab script to be used as template for participants proposals. This script performs all the steps for generating solutions for the Robot Vision challenge: features generation, training, classification and performance evaluation. Fig. 5 shows the same scene represented in a visual image and a point cloud data file. Training, validation and test sequences were acquired within two different buildings with similar room distribution structure. All the room and object categories included in the test sequence were previously seen during training and validation. No subtasks were considered, and therefore all participants have to prepare their submissions using the same single test sequence where the temporal continuity is not represented.

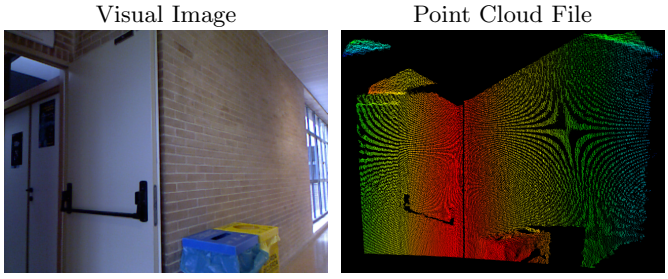


Fig. 5: Visual and 3D point cloud representation for a scene. Room class: corridor. List of objects: trash.

The 2014 dataset consists of three sequences (training, validation and test) of depth and visual images acquired within the following indoor environment: two department buildings at the University of Alicante, in Spain. Visual images were stored in PNG format while depth ones in PCD. Every image in the dataset is labelled with its corresponding room category and the list of eight objects to appear or not within it. The 10 room categories are: Corridor, Hall, ProfessorOffice, StudentOffice, TechnicalRoom, Toilet, Secretary, VisioConference, Warehouse and ElevatorArea. The 8 different objects are: Extinguisher, Phone, Chair, Printer, Urinal, Bookself, Trash and Fridge. The dataset has two labelled sequences used for training and validation with 5000 and 1500 images respectively. The unlabelled sequence used for test consists of 3000 different images. The frequency distribution for room categories and objects in the training, validation and test sequences are shown in Tables 6 and 7 respectively. Regarding the building used in the acquisition, all the 5000 training images were acquired in the building A. The validation sequence included 1000 images from building A but 500 new images from building B. Finally, all 3000 test images were acquired in building B.

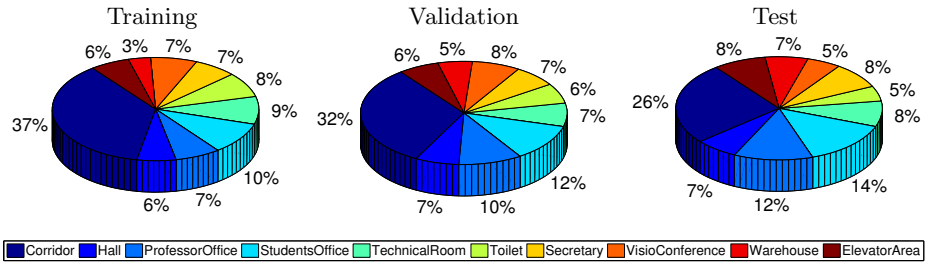


Fig. 6: Room distribution in training, validation and test sequences.

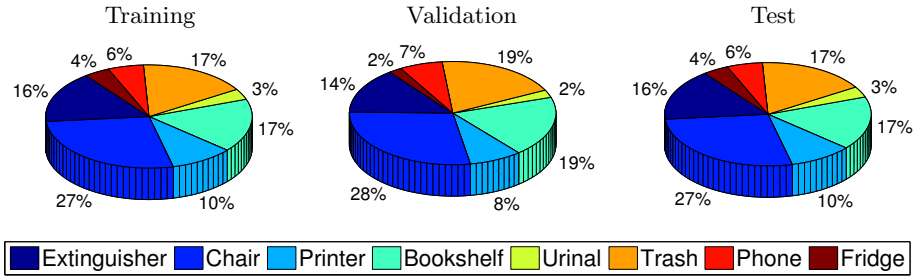


Fig. 7: Object distribution in training, validation and test sequences.

Participant submissions were compared and sorted according to the obtained score. Every submission consisted of the room category assigned to each test image and the corresponding list of the 8 detected/non-detected objects within the image. The number of times a specific object appears in an image is not relevant to compute the score. Participants are allowed to not classify rooms, in which case the score is not affected. The total score was computed using the rules shown in Table 3.

Table 3: Rules used to calculate the final score for a test frame

Room class/Category	
Room class/category correctly classified	+1.0 points
Room class/category wrongly classified	-0.5 points
Room class/category not classified	+0.0 points
Object Recognition	
For each object correctly detected (True Positive)	+1.0 points
For each object incorrectly detected (False Positive)	-0.25 points
For each object correctly detected as not present (True Negative)	+0.0 points
For each object incorrectly detected as not present (FalseNegative)	-0.25 points

Participation and Results In 2014, 28 participants registered to the Robot Vision task but only 4 submitted at least one run accounting for a total of 17 different runs. The scores obtained by all the submitted runs are shown in Fig. 8. The maximum score that could be achieved was 7004 (3000 from rooms and 4004 from objects) and the winner (NUDT) obtained a score of 4430,25 points.

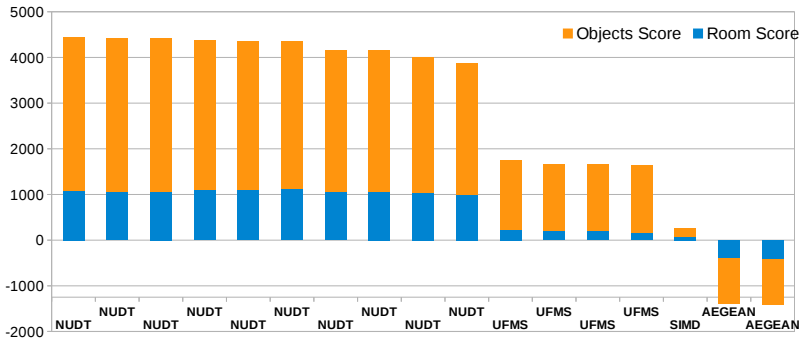


Fig. 8: Overall results of the runs submitted by the participant groups to the 2014 Robot Vision task.

The NUDT proposal [42] that ranked first followed a spatial pyramid matching approach [14] based on gradients and dense SIFT features. The classification was performed using a multi-class SVM following the one versus all strategy. The CPPP/UFMS proposal [4] also uses dense SIFT descriptors and the spatial pyramid approach. However, this approach is based on a k-nearest neighbor classifier. The SIMD proposal was generated by the organizers of the task using the proposed Matlab script (Depth+RGB histograms descriptors and SVM for classification).

In view of the results, we can conclude that room classification remains an open problem when generalization is requested. On the other hand, we should point out the high performance of the submissions when facing the object recognition problem. This can be explained because generalization is not needed to recognize a specific object within a scene. Namely, phones or chairs will always be recognized as their type (a phone or a chair respectively) independently from the scene where they are placed.

3 Conclusions

This paper presents an overview of the activities in the 2014 edition of the ImageCLEF lab. The sustained interest in the lab, witnessed by the important number of registration and the number of groups actually participating in the lab, make ImageCLEF an important resource in the image annotation 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 2015 edition of ImageCLEF.

References

1. Bosch, Anna ad Zisserman, A.: Image classification using random forests and ferns. In: Proc. CVPR (2007)
2. Caputo, B., Müller, H., Martinez-Gomez, J., Villegas, M., Acar, B., Patricia, N., Marvasti, N., Üsküdarlı, S., Paredes, R., Cazorla, M., Garcia-Varea, I., Morell, V.: ImageCLEF 2014: Overview and analysis of the results. In: CLEF proceedings. Lecture Notes in Computer Science, Springer Berlin Heidelberg (2014)
3. Caputo, B., Patricia, N.: Overview of the ImageCLEF 2014 Domain Adaptation Task. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes (2014)
4. de Carvalho Gomes, R., Correia Ribas, L., Antônio de Castro Junior, A., Nunes Gonçalves, W.: CPPP/UFMS at ImageCLEF 2014: Robot Vision Task. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes (2014)
5. Del Frate, F., Pacifici, F., Schiavon, G., Solimini, C.: Use of neural networks for automatic classification from high-resolution images. Geoscience and Remote Sensing, IEEE Transactions on 45(4), 800–809 (2007)
6. Feng, S., Manmatha, R., Lavrenko, V.: Multiple bernoulli relevance models for image and video annotation. In: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. vol. 2, pp. II–1002. IEEE (2004)

7. Friedl, M.A., Brodley, C.E.: Decision tree classification of land cover from remotely sensed data. Remote sensing of environment 61(3), 399–409 (1997)
8. Goh, K.S., Chang, E.Y., Li, B.: Using one-class and two-class svms for multiclass image annotation. Knowledge and Data Engineering, IEEE Transactions on 17(10), 1333–1346 (2005)
9. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: Proc. CVPR. Extended version considering its additional material
10. Jie, L., Tommasi, T., Caputo, B.: Multiclass transfer learning from unconstrained priors. In: Proc. ICCV (2011)
11. Kim, S., Park, S., Kim, M.: Image classification into object/non-object classes. In: Image and Video Retrieval, pp. 393–400. Springer (2004)
12. Ko, B.C., Lee, J., Nam, J.Y.: Automatic medical image annotation and keyword-based image retrieval using relevance feedback. Journal of Digital Imaging 25(4), 454–465 (2012)
13. Kökciyan, N., Türkay, R., Üsküdarlı, S., Yolum, P., Bakır, B., Acar, B.: Semantic Description of Liver CT Images: An Ontological Approach. IEEE Journal of Biomedical and Health Informatics (2014)
14. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. vol. 2, pp. 2169–2178. IEEE (2006)
15. Martinez-Gomez, J., Garcia-Varea, I., Caputo, B.: Overview of the imageclef 2012 robot vision task. In: CLEF (Online Working Notes/Labs/Workshop) (2012)
16. Martinez-Gomez, J., Garcia-Varea, I., Cazorla, M., Caputo, B.: Overview of the imageclef 2013 robot vision task. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes (2013)
17. Martinez-Gomez, J., Cazorla, M., Garcia-Varea, I., Morell, V.: Overview of the ImageCLEF 2014 Robot Vision Task. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes (2014)
18. Mueen, A., Zainuddin, R., Baba, M.S.: Automatic multilevel medical image annotation and retrieval. Journal of digital imaging 21(3), 290–295 (2008)
19. Muller, H., Clough, P., Deselaers, T., Caputo, B.: ImageCLEF: experimental evaluation in visual information retrieval. Springer (2010)
20. Park, S.B., Lee, J.W., Kim, S.K.: Content-based image classification using a neural network. Pattern Recognition Letters 25(3), 287–300 (2004)
21. Patricia, N., Caputo, B.: Learning to learn, from transfer learning to domain adaptation: a unifying perspective. In: Proc. CVPR (2014)
22. Pronobis, A., Caputo, B.: The robot vision task. In: Muller, H., Clough, P., Deselaers, T., Caputo, B. (eds.) ImageCLEF, The Information Retrieval Series, vol. 32, pp. 185–198. Springer Berlin Heidelberg (2010)
23. Pronobis, A., Christensen, H., Caputo, B.: Overview of the imageclef@ icpr 2010 robot vision track. Recognizing Patterns in Signals, Speech, Images and Videos pp. 171–179 (2010)
24. Qi, X., Han, Y.: Incorporating multiple svms for automatic image annotation. Pattern Recognition 40(2), 728–741 (2007)
25. Reshma, I.A., Ullah, M.Z., Aono, M.: KDEVIR at ImageCLEF 2014 Scalable Concept Image Annotation Task: Ontology based Automatic Image Annotation. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes. Sheffield, UK (September 15-18 2014)

26. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: Proc. ECCV (2010)
27. Sahbi, H.: CNRS - TELECOM ParisTech at ImageCLEF 2013 Scalable Concept Image Annotation Task: Winning Annotations with Context Dependent SVMs. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. Valencia, Spain (September 23-26 2013)
28. Sethi, I.K., Coman, I.L., Stan, D.: Mining association rules between low-level image features and high-level concepts. In: Aerospace/Defense Sensing, Simulation, and Controls. pp. 279–290. International Society for Optics and Photonics (2001)
29. Shi, R., Feng, H., Chua, T.S., Lee, C.H.: An adaptive image content representation and segmentation approach to automatic image annotation. In: Image and Video Retrieval, pp. 545–554. Springer (2004)
30. Tommasi, T., Caputo, B.: Frustratingly easy nbnn domain adaptation. In: Proc. ICCV (2013)
31. Tommasi, T., Quadrianto, N., Caputo, B., Lampert, C.: Beyond dataset bias: Multi-task unaligned shared knowledge transfer. In: Proc. ACCV (2012)
32. Tsikrika, T., de Herrera, A.S., M?ller, H.: Assessing the scholarly impact of imageclef. In: Cross Language Evaluation Forum (CLEF 2011). Lecture Notes in Computer Science (LNCS), Springer (2011)
33. Ünay, D., Soldea, O., Akyüz, S., Çetin, M., Erçil, A.: Medical image retrieval and automatic annotation: Vpa-sabanci at imageclef 2009. The Cross-Language Evaluation Forum (CLEF) (2009)
34. Vailaya, A., Figueiredo, M.A., Jain, A.K., Zhang, H.J.: Image classification for content-based indexing. Image Processing, IEEE Transactions on 10(1), 117–130 (2001)
35. Villegas, M., Paredes, R.: Overview of the ImageCLEF 2012 Scalable Web Image Annotation Task. In: Forner, P., Karlgren, J., Womser-Hacker, C. (eds.) CLEF 2012 Evaluation Labs and Workshop, Online Working Notes. Rome, Italy (September 17-20 2012), http://mvillegas.info/pub/Villegas12_CLEF_Annotation-Overview.pdf
36. Villegas, M., Paredes, R.: Overview of the ImageCLEF 2014 Scalable Concept Image Annotation Task. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes. Sheffield, UK (September 15-18 2014), http://mvillegas.info/pub/Villegas14_CLEF_Annotation-Overview.pdf
37. 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 (September 23-26 2013), http://mvillegas.info/pub/Villegas13_CLEF_Annotation-Overview.pdf
38. Villena Román, J., González Cristóbal, J.C., Goñi Menoyo, J.M., Martínez Fernández, J.L.: MIRACLE's naive approach to medical images annotation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 28(7), 1088–1099 (2005)
39. Wong, R.C., Leung, C.H.: Automatic semantic annotation of real-world web images. Pattern Analysis and Machine Intelligence, IEEE Transactions on 30(11), 1933–1944 (2008)
40. Yang, C., Dong, M., Fotouhi, F.: Image content annotation using bayesian framework and complement components analysis. In: Image Processing, 2005. ICIP 2005. IEEE International Conference on. vol. 1, pp. I–1193. IEEE (2005)
41. Yılmaz, K.Y., Cemgil, A.T., Simsekli, U.: Generalised coupled tensor factorisation. In: Advances in Neural Information Processing Systems. pp. 2151–2159 (2011)

42. Zhang, Y., Qin, J., Chen, F., Hu, D.: NUDTs Participation in ImageCLEF Robot Vision Challenge 2014. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes (2014)