General Overview of ImageCLEF at the CLEF 2015 Labs

Mauricio Villegas[1], Henning Müller[2], Andrew Gilbert[3], Luca Piras[4], Josiah Wang[5], Krystian Mikolajczyk[3], Alba G. Seco de Herrera[2], Stefano Bromuri[2], M. Ashraful Amin[6], Mahmood Kazi Mohammed[7], Burak Acar[8], Suzan Uskudarli[8], Neda B. Marvasti[8], José F. Aldana[9], and María del Mar Roldán García[9]

[1] Universitat Politècnica de València, Spain
[2] University of Applied Sciences Western Switzerland (HES-SO), Switzerland
[3] University of Surrey, UK
[4] University of Cagliari, Italy
[5] University of Sheffield, UK
[6] Independent University, Bangladesh
[7] Sir Salimullah Medical College, Bangladesh
[8] Bogazici University, Turkey
[9] University of Malaga, Malaga, Spain

Abstract. This paper presents an overview of the ImageCLEF 2015 evaluation campaign, an event that was organized as part of the CLEF labs 2015. ImageCLEF is an ongoing initiative that promotes the evaluation of technologies for annotation, indexing and retrieval for providing information access to databases of images in various usage scenarios and domains. In 2015, the 13th edition of ImageCLEF, four main tasks were proposed: 1) automatic concept annotation, localization and sentence description generation for general images; 2) identification, multi-label classification and separation of compound figures from biomedical literature; 3) clustering of x-rays from all over the body; and 4) prediction of missing radiological annotations in reports of liver CT images. The x-ray task was the only fully novel task this year, although the other three tasks introduced modifications to keep up relevancy of the proposed challenges. The participation was considerably positive in this edition of the lab, receiving almost twice the number of submitted working notes papers as compared to previous years.

1 Introduction

In the current age of the Internet and the proliferation of increasingly cheaper devices to capture, amongst others, visual information, developing technologies for the storage of this ever growing body of information and providing means to access these huge databases is and will be a requirement. As part of this development, it is of great importance to organise campaigns for evaluating the emerging problems and for fairly comparing the proposed techniques for solving them. Motivated by this, now in its 13th edition, ImageCLEF has been an

initiative aimed at evaluating multilingual or language independent annotation and retrieval of images [14]. The main goal of ImageCLEF is to support the advancement of the field of visual media analysis, classification, annotation, indexing and retrieval, by proposing novel challenges and developing the necessary infrastructure for the evaluation of visual systems operating in different contexts and providing reusable resources for benchmarking.

To meet its objectives, ImageCLEF organises tasks that benchmark the annotation, classification and retrieval of diverse images such as the heterogeneous images found on web pages as well as imagery used in specialised fields such as medicine. These tasks aim to support and promote research that addresses key challenges in the field. ImageCLEF has had a significant influence [20] on the visual information retrieval field by benchmarking various retrieval, classification 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 [19]. One offspring of ImageCLEF is LifeCLEF [9] that includes besides images of leaves (a former ImageCLEF task) now also videos of fish that need to be identified and sounds of birds, making it a real multimedia retrieval task. Another CLEF lab linked to ImageCLEF is CLEFeHealth [6] that deals with information retrieval from health-related documents. Also the eHealth lab is coordinated in close collaboration with ImageCLEF, as there is an overlap with the medical task.

This paper presents a general overview of the ImageCLEF 2015 evaluation campaign[10], which as usual was an event organised as part of the CLEF labs[11]. The remainder of this paper is organised as follows. Section 2 starts with a general description of the 2015 edition of ImageCLEF, commenting about the overall organisation and participation in the lab. Followed by this are subsections dedicated to the four main tasks that were organised this year, Section 2.1 for the image annotation task, Section 2.2 for the medical classification task, Section 2.3 for the medical clustering task and Section 2.4 for the liver CT annotation task. Finally, the paper concludes with Section 3 giving an overall discussion, and pointing towards the challenges ahead and possible new directions for ImageCLEF 2016.

2 ImageCLEF 2015: the Tasks, the Data and the Participation

The 2015 edition of ImageCLEF consisted of four main tasks that covered challenges in diverse fields and usage scenarios. Similar to the 2014 edition [2], all of the tasks addressed topics related to processing the images in order to auto-

[10] http://imageclef.org/2015/
[11] http://clef2015.clef-initiative.eu/

matically assign meta-data to them, not directly evaluating retrieval, but techniques that produce valuable annotations that can be used for subsequent image database indexing, mining or analysis. The four tasks organised were the following:

- **Image Annotation:** aims at developing systems for automatic annotation of concepts, their localization within the image, and generation of sentences describing the image content in a natural language.
- **Medical Classification:** addresses the identification, multi-label classification and separation of compound figures commonly found in the biomedical literature.
- **Medical Clustering:** is a task of which the objective is to cluster a dataset of x-rays from all over the body in order to group them according to the body part that is visible.
- **Liver CT Annotation:** has as goal the prediction of missing radiological annotations in structured radiology reports of liver CT images based on a new ontology of liver cases LiCO.

In order to participate in the evaluation campaign, the groups first had to register either on the CLEF website or from the ImageCLEF website. To actually get access to the datasets, the participants were required to submit by email a signed End User Agreement (EUA). Table 1 presents figures that summarize the participation in ImageCLEF 2015, including the number of registrations and number of signed EUAs, indicated both per task and for the overall lab. The table also includes the number of groups that submitted results (also called runs) and the ones that submitted a working notes paper describing the techniques used.

The number of registrations could be interpreted as the initial interest that the community has for the evaluation. However, it is a bit misleading because several people from the same institution might register, even though in the end they would count as a single group participation. The EUA explicitly requires all groups that get access to the data to participate. Unfortunately, the percentage of groups that take part is relatively small. Nevertheless, as observed in studies of scholarly impact [19], in subsequent years the datasets and challenges provided by ImageCLEF do get used quite often, which in part is due to the researchers that for some reason were unable to participate in the original event.

A very positive result for the 2015 edition of ImageCLEF was the number of working notes paper submissions, which can be considered the most important outcome, since this indicates the number of evaluated techniques that get properly reported in the literature. In total 25 papers were submitted, which in comparison to the previous two years (11 papers for 2013 and 13 papers for 2014), the participation has almost doubled.

The following four subsections are dedicated to each of the tasks. Only a short overview is reported, including general objectives, description of the tasks and datasets and a short summary of the results. For more detais please refer to the corresponding task overview papers [5,8,1,13].

Table 1: Key figures of participation in ImageCLEF 2015.

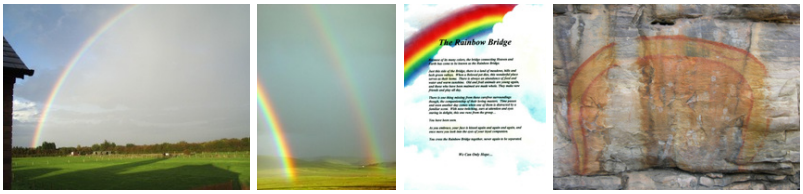
Task	Online registrations	Signed EUA	Groups that subm. results	Submitted working notes
Image Annotation	92	47	14	11
Medical Classification	79	34	8	7
Medical Clustering	72	36	8	6
Liver CT Annotation	51	27	1	1
Overall	148	72	31*	25*

* Total for all tasks, not unique groups.

2.1 The Image Annotation Task

Every day, users struggle with the ever-increasing quantity of data available to them. Trying to find "that" photo they took on holiday last year, the image on Google of their favourite actress or band, or the images of the news article someone mentioned at work. There is a large number of images that can be cheaply found and gathered from the Internet. However, more valuable is mixed modality data, for example, web pages containing both images and text. A large amount of information about the image is present on these web pages and vice-versa. However, the relationship between the surrounding text and images varies greatly, with much of the text being redundant and/or unrelated. Despite the obvious benefits of using such information in automatic learning, the very weak supervision it provides means that it remains a challenging problem. The scalable concept annotation, localization and sentence generation task aims to develop techniques to allow computers to reliably describe images, localize the different concepts depicted in the images and generate a description of the scene. This year the task was split into three related subtasks using a single mixed modality data source of 500,000 web page items.

Past Editions The Scalable Concept annotation, localization and sentence generation 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 2015 Scalable Concept Image Annotation task



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

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

Fig. 1: Examples of images retrieved by a commercial image search engine.

is its fourth edition, having been organized in 2012 [21], 2013 [23] and 2014 [22]. In light of recent interest in annotating images beyond just concept labels, we introduced two new subtasks this year where participants developed systems to describe an image with a textual description of the visual content depicted in the image.

Objective and Task for the 2015 Edition Image concept annotation, localization and natural sentence generation generally has relied on training data that has been manually, and thus reliably annotated, an expensive and laborious endeavour 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 are very noisy, but overall these data contain useful information that can be exploited to develop annotation systems. Figure 1 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.

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, localizing, and generating natural sentences from practically any topic. There were 3 sub tasks available to participants, which all use the common 500,000 web pages of images and text training data. Unlike previous years the test set was also the training set.

1. **SubTask 1:** The image annotation task continues in the same line of past years. The objective required the participants to develop a system that receives as input an image and produces as output a prediction of which concepts are present in that image, selected from a predefined list of concepts and starting this year, where they are located within the image.
2. **SubTask 2:** *Clean track*. In light of recent interest in annotating images beyond just concept labels, this subtask required the participants to describe images with a textual description of the visual content depicted in the image. It is thought of as an extension of SubTask 1. Aimed primarily at those interested only in the Natural Language Generation aspects of the subtask, therefore a gold standard input (bounding boxes labelled with concepts) was provided to develop systems that generate sentence, (natural language based) descriptions based on these gold standard annotations as input.
3. **SubTask 3:** *Noisy Track* This track was geared towards participants interested in developing systems that generated textual descriptions directly from images, e.g. by using visual detectors to identify concepts and generating textual descriptions from the detected concepts. This had a large overlap with sub task 1.

The concepts this year were chosen to be visual objects that are localizable and that are useful for generating textual descriptions of visual content of images. They include animate objects such as person, dogs and cats, inanimate objects such as houses, cars and balls, and scenes such as city, sea and mountains. The concepts were mined from the texts of our large database of image-webpage pairs. Nouns that are subjects or objects of sentences are extracted and mapped onto WordNet synsets. These were then filtered to 'natural', basic-level categories ('dog' rather than a 'yorkshire terrier'), based on the WordNet hierarchy and heuristics from a large-scale text corpora. The final list of concepts were manually shortlisted by the organizers such that they were (i) visually concrete and localizable; (ii) suitable for use in image descriptions; (iii) at a suitable 'every day' level of specificity that were neither too general nor too specific.

The data used in this task was similar to the one from last year [22]. The training and test set was composed of 500,000 samples each of which included: the raw image, pre-computed visual features and textual features. These training images were obtained from the web by querying popular image search engines. The development and sub task 1 and 3 test sets were both taken from the "training set" and had 1,979 and 3,070 samples, and the clean sub task 2 track had 500 and 450 samples. For further details, please refer to the task overview paper [5].

Participation and Results This year 14 groups participated in the task, submitting a total of 122 runs across the 3 sub tasks and 11 of the participants also submitted working notes papers. Further details on the specific sub tasks is shown below. Sub task 1 was well received despite the additional requirement of labeling and localizing all 500,000 images. The ground truth used for the evaluation of the approaches used an unknown small subset of the 500,000 images.

Table 2: Sub task 1 results.

Group	**0% Overlap**	**50% Overlap**
SMIVA	0.79	0.66
IVANLPR	0.64	0.51
Multimedia Comp Lab	0.62	0.50
RUC	0.61	0.50
CEA	0.45	0.29
Kdevir	0.39	0.23
ISIA	0.25	0.17
CNRS-TPT	0.31	0.17
IRIP-iCC	0.61	0.12
UAIC	0.27	0.06
MLVISP6	0.06	0.02
REGIM	0.03	0.02
Lip6	0.04	0.01

Localization of Sub task 1 was evaluated using the PASCAL style metric of intersection over union (IoU), the area of intersection between the foreground in the output segmentation and the foreground in the ground-truth segmentation, divided by the area of their union. The final results are presented in table 2 in terms of mean average precision (MAP) over all images of all concepts, with both 0% overlap (i.e. no localization) and 50% overlap. It can be seen that four groups have achieved over 50 MAP across the evaluation set with 50% overlap with the ground-truth. This seems an excellent result given the challenging nature of the images used and the wide range of concepts provided. SMIVA used a deep learning framework with additional annotated data, while IVANLPR implemented a two-stage process, initially classifying at an image level with an SVM classifier, and then applying deep learning feature classification to provide localization. The Multimedia Comp Lab gathered high-quality training examples from the Web, then per concept, an ensemble of linear SVMs is trained by Negative Bootstrap, with CNN features as image representation. A shortcoming of the overall challenge however is the difficulty of ensuring the ground truth has 100% of concepts labelled, thus allowing a recall measure to be used. With the current crowd source based hand labelling of the ground truth it was found not to achieve this and so a recall measure isn't evaluated.

The pilot sub tasks on sentence generation received a reasonably good amount of participation, with two groups participating in sub task 2 and four in sub task 3. We observed a variety of approaches used to tackle these sub tasks, including top-down approaches, deep learning methods and joint image-text retrieval. Both sub tasks were evaluated using the Meteor evaluation metric [3]. We have also pioneered an additional fine-grained metric for sub task 2, which is the average F1 score across 450 test images on how well the sentence generation system selects the correct concepts to be described against gold standard image descriptions. Table 3 shows the results of the best run for each participant. For sub task 2, both

Table 3: Sub task 2 and 3 results

Group	Sub task 2		Sub task 3
	F1 score	**Meteor**	**Meteor**
ISIA	–	–	0.1687 ± 0.0852
MindLab	–	–	0.1403 ± 0.0564
RUC	0.5310 ± 0.2327	0.2393 ± 0.0865	0.1875 ± 0.0831
UAIC	0.5030 ± 0.1775	0.2097 ± 0.0660	0.0813 ± 0.0513

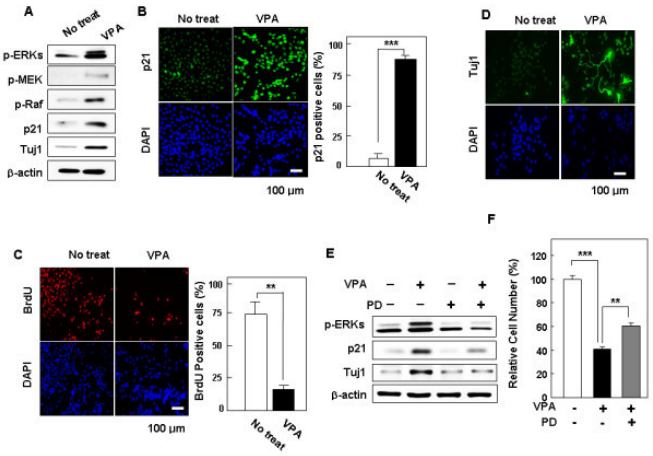
groups achieved comparable results for both evaluation metrics. For sub task 3, three groups achieved Meteor scores of over 0.10. The encouraging participation rates and promising results in these pilot sub tasks are sufficient motivations for the sub tasks to be included in future editions of this challenge.

For the complete results and a more detailed analysis, the reader should refer to the task overview paper [5].

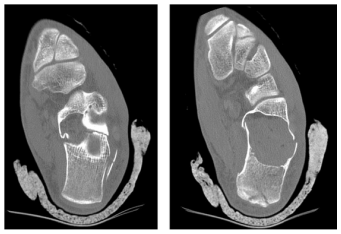
2.2 The Medical Classification Task

This task is motivated by the fact that an estimated 40% of the figures in PubMed Central are compound figures (images consisting of several sub figures) [7]. Examples of compound figures can be seen in Figure 2. When data of articles are made available digitally, often the compound images are not available separated but made available in a single block. Information retrieval systems for images should be capable of distinguishing the parts of compound figures that are potentially relevant to a given query. A major step for 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 via their visual content. More information about this task can also be found in [8].

Past Editions The medical image retrieval and classification task has been held every year at ImageCLEF since 2004, apart from 2014 [10]. The goal has been to promote biomedical image retrieval by combining text and images for more effective multimodal retrieval. It is also possible to use image modality classification to filter retrieval result lists or rerank them to improve and focus the retrieval. Therefore, in 2010, a modality classification task was introduced. The classification hierarchy has evolved over the years to an improved ad hoc hierarchy with 31 classes in 2012. It includes sections of diagnostic images, generic biomedical illustrations and compound or multipane images [15]. In 2013, the same hierarchy as in ImageCLEF 2012 was used. However, a larger number of compound figures than in ImageCLEF 2012 were provided in the training and test sets. The distribution of compound vs. non-compound figures corresponds



(a) Mixed modalities in a single figure.



(b) Two images from the same modality in a single figure.

Fig. 2: Examples of compound figures found in the biomedical literature.

to that in the PubMed Central data set[12] that the traing and test set are part of.

Making the content of the compound figures accessible for targeted search can improve retrieval accuracy. For this reason the detection of compound figures and their separation into subfigures was introduced in 2013 [7].

Objective and Task for the 2015 Edition In 2015, the task focused only on the compound figures and not on potential retrieval steps after the compound figure analyis. There are four subtasks in 2015:

– *Compound figure detection* – Compound figure identification is a required first step to separate compound images from images only containing a single type of content. Therefore, the goal of this subtask is to identify whether a figure is a compound figure or not. The task makes training data available containing compound and non compound figures from the biomedical literature that are labelled and then a test set with similar images.

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

– *Multi-label classification* – Characterization of compound figures is difficult, as they may contain subfigures from various imaging modalities or image types. This task aims to label each compound figure with each of the modalities (of the 31 classes of a defined hierarchy) of the subfigures contained without knowing where the separation lines are. Differently from previous years, in which we focused on separating the subfigures to classify them on their own, we decided to consider the entire compound figure as having multiple labels simultaneously. We expect that this approach may help identifying the classes in the subfigures by taking into consideration the relationships occurring among the classes during the training of the model.
– *Figure separation* – This subtask was first introduced in 2013. The task makes available training data with separation labels of the figures and then a test data set where the labels are made available after the submission of the results. In 2015, a larger number of compound figures was distributed compared to the previous subtask.
– *Subfigure classification* – Similar to the modality classification task organized in 2011-2013 this subtask aims to classify images into the 31 classes of the hierarchy. The images are the subfigures extracted from the compound figures distributed for the figure separation subtask.

Participation and Results Over seventy groups registered for the medical classification tasks and 8 groups submitted at least one run and a working notes paper. 40 runs were submitted in this task in total.

The FHDO Biomedical Computer Science Group [16] obtained best results on the compound figure detection and subfigure classification subtasks achieving 85% and 68% of accuracy respectively. In the multi-label classification, the MindLab group[13] obtained the best result with a Hamming loss of 0.05, while the IIS [17] group obtained a very close result to MindLab with a 0.671 in terms of Hamming loss. Finally, the National Library of Medicine (NLM) [18] submitted the best run in the figure separation subtask achieving an accuracy of 85%.

A more detailed analysis of the medical classification tasks is presented in the task overview paper of the working notes [8].

2.3 The Medical Clustering Task

At our research centre we are developing a diagnostic imaging teaching and learning system for medical students of Bangladesh [4]. As part of this project, a large collection of digital x-ray images was created from data obtained at a local hospital. These data are being used to build our teaching and learning system. However, during this development process the archiving and retrieving of x-ray images from a large database was found to be a challenging problem. Thus we decided to open up this challenge to the community by organising it as an evaluation under the framework of ImageCLEF.

[13] https://sites.google.com/a/unal.edu.co/mindlab

Fig. 3: Example images from the training data set.

Objective and Task for the 2015 Edition The primary objective of this task is to group digital x-ray images into four major clusters: head-neck, upper-limb, body, and lower-limb. The secondary goal of this task is to partition the initial clusters into sub-clusters, for example the upper-limb cluster can be farther divided into: Clavicle, Scapula, Humerus, Radius, Ulna, and Hand. However, due to the time constraint and difficulty level of the task, this year we decided to go only with the primary objective.

All together there are 500 digital x-ray images in the training dataset, of which 100 are from each of the four desired clusters: head-neck, upper-limb, body, and lower-limb, and the remaining 100 are true negative images that are taken by the same digital x-ray camera for calibration purpose. Some example images are given in Figure 3. A point to be noted about the assigned classification information for an x-ray image is that a single image can belong to multiple classes. For example, if a full body x-ray image for a child is given, then the associated classes are: head-neck, upper-limb, body, and lower-limb and the associated output is a 4-bit string that should have the value [1 1 1 1]. 250 test images were made available to the participants to check the performance of their system. At this moment, the task organizers have made the 750 samples available. They intend to make all 5000 images available in high resolution as '.dcm' format for non-commercial use at[14] soon after the CLEF 2015 conference.

Participation and Results 71 groups from all 6 continents of the world participated in the initial level and acquired data from the ImageCLEF website. Though it is primarily a European event, 15 groups from EU, 14 from North America, 6 from Australia and 29 from Asia participated in the initial event. Among all EU countries there were 5 German groups and in Asia China had 5 groups which is highest in their region. Finally, participants were given the test data and a month time to submit their results on the test data. Only, 8 groups submitted their final results. There were 2 submissions from Australia, 2 from USA, 1 from each of the countries Republic of Korea, Israel, China and none from the EU. One group withdrew their runs (submitted results) as their method was semi-automatic. 7 groups submitted 29 runs and the best results

[14] http://www.cvcrbd.org

for each group are selected and provided in Table 4. Finally, 6 groups were able to submit working note papers describing the methods used to implement their x-ray clustering system.

To solve this multiclass classification problem of grouping digital x-ray image in to four clusters, participants have taken different approaches. For feature extraction they utilized: Intensity Histogram (IH), Gradient Magnitude Histogram (GM), Shape Descriptor Histogram (SD), Curvature Descriptor Histogram (CD), Histogram of Oriented Gradient (HOG), Local Binary Pattern (LBP), Color Layout Descriptor (CLD), Edge Histogram Descriptor (EHD) from MPEG-7 standard, Color and Edge Direction Descriptor (CEDD), Fuzzy Color and Texture Histogram (FCTH), Tamura texture descriptor, Gabor texture feature, primitive length texture features, edge frequency texture features, autocorrelation texture features, Bag of Visual Words (BoVW), Scale invariant feature transform (SIFT), Speeded up robust features (SURF), Binary robust independent elementary features Brief (BRIEF), Oriented fast and rotated BRIEF (ORB), Multiscale LBP Histogram with Spatial Pyramid, Sparse Coding with Max-pooling and Spatial Pyramid, Fisher Kernel Feature Coding, Global mean of rows and columns, Local Mean of rows and columns, and Gray Level Co-occurrence Matrix (GLCM).

Classification was performed using Backpropagation Neural Networks (BPNN), Logistic Regression (LR), K Nearest Neighbors (KNN) and Deep Belief Network (DBN), Convolution Neural networks (CNN), Decision Tree, Support Vector Machine (RBF Kernel, Poly kernel, Normalized Ploy kernel and Puk kernel), Random Forest, Logistic Model Tree (LMT), Naive Bayesian, and Ensemble Neural Network.

Because one input can belong to multiple classes, we have tested the performance based on three different methods. The most conventional one is the Hamming similarity calculation. However, a stricter version of classification accuracy checking is also used, that we are calling exact matching, which basically checks, for a given input how many of its multiple possible classes are correctly identified. We also checked the accuracies using another method that we call any match. For an input image if the predicted class matches with any of the actual class of that image then it is considered a correct classification. Best result for exact match was 0.752; for any match was 0.864; and for Hamming similarity was 0.895. Final score for all seven groups is provided in Table 4.

It is very likely that participants are using similar feature extraction and classification techniques. It is accepted that some features are used by most of the participants, those are the so-called state of the art techniques. However, for this problem of clustering x-ray images into 4 clusters 6 participants have employed 27 different image feature extraction techniques. Different characteristics of the feature extractors are revealed. One interesting observation is that while exploring the famous HoG features one group claims it has poor discriminating capacity and, on the other hand, another group is providing an accuracy above 0.90 using HoG features. Another interesting observation is that, even though, x-ray images are gray scale, color features like CEDD, FCTH show good discriminating

Table 4: Final Results of the Digital X-Ray Image clustering task.

Group	Exact Match Score	Any Match Score	Hamming Similarity
IBM MMAFL	0.752	0.864	0.863
SNUMedInfo	0.709	0.856	0.895
AmrZEGY	0.646	0.780	0.868
NLM	0.613	0.740	0.849
CASMIP	0.606	0.732	0.843
BMET	0.497	0.596	0.816
db Lab	0.219	0.264	0.664

ability [1]. Most interesting yet obvious observation is the use of Convolution Neural Network (CNN). Recently, CNN were made popular by GoogLeNet. Out of seven, four groups used or experimented with Neural Networks. It is good news for the neural network researchers. We believe people have already started (rather restarted) to explore enormous ability of CNN and other computational learners other than SVM's.

2.4 The Liver CT Annotation Task

Medical and more specifically radiological databases present challenges due to the exponential increase in data volumes. Radiological images contain a rich source of meta-data associated with the images. A significant part of the medical image analysis is based on the subtle differences between a set of similar images, such as abdominal CT images. In a conventional setting, these critical differences, such as the parenchyma texture of a liver, are manually observed by experts and are translated into the medical vocabulary. Domain-specific radiological structured-reporting is useful in accurately reflecting the interpretation of medical images. Such reporting can improve the clinical workflow by means of facilitating standardized reports as well as boost the performance of search and retrieval from radiological databases for the purpose of comparative diagnosis, medical education, etc. Despite its advantages, an expert annotation is a labour intensive task, which can be performed by qualified individuals only and must be consistent among different individuals, sites, countries, etc. Computer-aided automatic annotation is a challenging task, which facilitates filling in a structured radiology report. Several standard terminologies are being developed/used for medical annotation, such as SNOMED-CT (Systemized Nomenclature in Medicine), RadLex (Radiology Lexicon), NCBO, UMLS (Unified Medical Language System), LOINC (Logical Observation Identifiers Names and Codes), etc. An annotation is performed via a high-level processing of the medical evidence derived from the images. Hence, a key challenge in expert annotation is to translate computer generated objective low-level image observations (CoG) to high level semantic descriptions (i.e., annotations) that comply with a standard terminology of choice. The "Liver CT Annotation Task", aims at filling structured reports by facilitating the computer aided annotation of liver CT images.

Table 5: Results of the runs of Liver CT annotation task.

Run	Completeness	Accuracy	Total Score
Run1	0.990909	0.825688	0.904534
Run2	0.990909	0.822630	0.902857
Run3	0.990909	0.836391	0.910378

Past Edition The Liver CT Annotation task was introduced for the first time in ImageCLEF 2014 [12], which focused on the annotation of the liver CT images and filling structured reports generated using the ontology called ONLIRA (ONtology of Liver for Radiology) [11]. ONLIRA describes the imaging observations of the liver itself as well as vessel and lesions inside. Every term in the given structured report is defined by an ontology property (object/data).

Objective and Task for the 2015 Edition In 2015 [13], the ontology was enriched by adding patient and study level information to ONLIRA. The new ontology is called LiCO (Liver Case Ontology)[15]. Patient level contains general information about the current patient, which includes name, age, gender, regular drugs, surgeries, and diseases. Study level consists of nonregular drugs, different diagnosis, physical examination, and laboratory results. The participants were given a training set of 50 cropped liver CT images together with the liver masks, and a bounding box defining the lesion area, a set of semantic annotations regarding the patient, study and imaging observations generated automatically from structures reports based on LiCO. Imaging observations contain the liver, vessels and one selected lesion. The semantic features were generated by an expert radiologist as part of the CaReRa[15] (Case Retrieval in Radiology) project, using the open source LiCO. The test set had 10 cases, with all types of data available in the training set except the semantic features in RDF format. The participants were asked to estimate the missing 65 imaging observations (UsE features). They were allowed to use any feature extraction method to generate low-level imaging observations from the CT images. The evaluation was based on the completeness (defined as the percentage of all 65 UsE features that were estimated) and accuracy (defined as the percentage of the estimated UsE features that were correct), and geometric mean of which was used as the total Score. Ideally, all metrics are 1.00.

Participation and Results In 2015, there was 1 participant from Tlemcen University, who submitted 3 runs and 1 working note paper. Table 5 lists the results of all runs submitted. It can be seen that the third run outperforms the other two. Table 6 compares the results of different runs in predicting different groups of UsE features. We divide UsE features into 5 groups: liver, vessels and

[15] http://www.vavlab.ee.boun.edu.tr/pages.php?p=research/CARERA/carera.html

Table 6: Total score of the runs of Liver CT annotation task for different groups of features.

Group	**Run1**	**Run2**	**Run3**
Liver	0.925	0.925	0.925
Vessel	1.000	1.000	1.000
LesionArea	0.730	0.746	0.753
LesionLesion	0.470	0.470	0.480
LesionComponent	0.870	0.844	0.889

three lesion groups with area, lesion and component concepts. Results show that all methods have predicted the vessel UsE features completely. Also all runs have the same performance over liver features. The only difference is on lesion specified features, in which the third run outperforms the other. The first run is using a random forest classifier with liver texture and shape features, the second run is performed with the same method as the first run, except it employs the texture and shape features of the lesion. The third run is completed using the specific signature of the liver, which is done in 2D space on the slice located in the centre of the lesion. First, in order to forbid the imaging inconsistencies, they normalized the image into a rectangular block with constant dimensions, which is then divided into small blocks. Then, the 1D Log-Gabor filter is applied to each block and the dominant phase data is selected and quantized to 4 levels to encode the pattern of the liver. Finally, The similarity is calculated by Hamming distance and majority voting is then employed to assign the annotation.

For more details on the task and the results, the reader should refer to the task overview paper [13].

3 Conclusions

This paper presents a general overview of the activities and outcomes of the 2015 edition of the ImageCLEF evaluation campaign. Four main tasks were organised covering challenges in: automatic concept annotation, localization and sentence description generation for web images; identification, multi-label classification and separation of compound figures from biomedical literature; clustering of x-rays from all over the body; and prediction of missing radiological annotations in reports of liver CT images.

The interest in the lab was outstanding, receiving signed End User Agreements requesting access to the datasets from over seventy groups world wide. The participation in terms of submission of results was also quite satisfactory, receiving system runs from about thirty groups. In total 25 working notes papers were submitted describing the systems that were evaluated in all tasks, this being almost double than the previous years.

Even though the x-ray clustering task was in its first edition, several groups showed interest and submitted results. On the other hand, the Liver CT task in

its second edition had a lower than expected participation. In part this can be due to the difficulty of the problem, although it is possible that there was not enough advertising or the audience targeted was not fully appropriate. For the next editions of ImageCLEF a greater effort must be made to assure that all tasks are well advertised so that they all have good participation.

The other two tasks that have run for several years both introduced important modifications that seemed to be heading in the right directions. The compound figure task addressed all aspects of dealing with compound figures in the literature and the participants obtained good performances. On the other hand, the image annotation task introduced the requirements of locating the concepts within the image and generating a natural language description. The added difficulty did not hinder the participation, in fact it can be said that there was a renewed interest. The sentence description generation had fewer participants although it has a great potential so it should continue in future editions.

Acknowledgements

The general coordination has been supported by the European Science Foundation (ESF) through the research networking programme Evaluating Information Access Systems (ELIAS). The Scalable Concept image annotation, localization and sentence generation task is co-organized by the ViSen consortium under the EU CHIST-ERA D2K Programme, supported by EPSRC Grants EP/K01904X/1 and EP/K019082/1. The medical clustering task at the 2015 ImageCLEF is supported by the Independent University Bangladesh and European Science Foundation's (ESF) Research Networking Programmes (RNPs). The Liver CT Annotation task is supported by TÜBİTAK Grant # 110E264 (CaReRa project) and in part by COST Action IC1302 (KEYSTONE).

References

1. Amin, M.A., Mohammed, M.K.: Overview of the ImageCLEF 2015 medical clustering task. In: CLEF2015 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Toulouse, France (September 8-11 2015)
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: Information Access Evaluation. Multilinguality, Multimodality, and Interaction, Lecture Notes in Computer Science, vol. 8685, pp. 192–211. Springer International Publishing (2014), doi:10.1007/978-3-319-11382-1_18
3. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the EACL 2014 Workshop on Statistical Machine Translation (2014)
4. Faruque, M.S.S., Banik, S., Mohammed, M.K., Hasan, M., Amin, M.A.: Teaching and learning system for diagnostic imaging phase i: X-ray image analysis and retrieval. In: Proceedings of the 6th International Conference on Computer Supported Education (2015)

5. Gilbert, A., Piras, L., Wang, J., Yan, F., Dellandrea, E., Gaizauskas, R., Villegas, M., Mikolajczyk, K.: Overview of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task. In: CLEF2015 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Toulouse, France (September 8-11 2015)
6. Goeuriot, L., Kelly, L., Li, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Jones, G., Müller, H.: Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval clef ehealth overview. In: CLEF Proceedings. Springer LNCS (2014)
7. 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) (September 2013), http://ceur-ws.org/Vol-1179/CLEF2013wn-ImageCLEF-SecoDeHerreraEt2013b.pdf
8. García Seco de Herrera, A., Müller, H., Bromuri, S.: Overview of the ImageCLEF 2015 medical classification task. In: Working Notes of CLEF 2015 (Cross Language Evaluation Forum). CEUR Workshop Proceedings, CEUR-WS.org (September 2015)
9. Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, Willem-Pier amd Planque, R., Rauber, A., Fisher, R., Müller, H.: Lifeclef 2014: Multimedia life species identification challenges. In: CLEF Proceedings. Springer LNCS (2014)
10. Kalpathy-Cramer, J., García Seco de Herrera, A., Demner-Fushman, D., Antani, S., Bedrick, S., Müller, H.: Evaluating performance of biomedical image retrieval systems –an overview of the medical image retrieval task at imageclef 2004-2014. Computerized Medical Imaging and Graphics 39, 55 –61 (2015), doi:10.1016/j.compmedimag.2014.03.004
11. Kokciyan, N., Turkay, R., Uskudarli, S., Yolum, P., Bakir, B., Acar, B.: Semantic description of liver ct images: An ontological approach. Biomedical and Health Informatics, IEEE Journal of PP(99), 1–1 (2014), doi:10.1109/JBHI.2014.2298880
12. Marvasti, N., Kökciyan, N., Türkay, R., Yazıcı, A., Yolum, P., Üsküdarlı, S., Acar, B.: ImageCLEF Liver CT Image Annotation Task 2014. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes (2014), http://ceur-ws.org/Vol-1180/CLEF2014wn-Image-MarvastiEt2014.pdf
13. Marvasti, N.B., del Mar Roldán García, M., Uskudarli, S., Aldana, J.F., Acar, B.: Overview of the ImageCLEF 2015 liver CT annotation task. In: CLEF2015 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Toulouse, France (September 8-11 2015)
14. Müller, H., Clough, P., Deselaers, T., Caputo, B.: ImageCLEF: experimental evaluation in visual information retrieval. Springer-Verlag Berlin Heidelberg (2010), doi:10.1007/978-3-642-15181-1
15. 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) (September 2012), http://ceur-ws.org/Vol-1178/CLEF2012wn-ImageCLEF-MullerEt2012.pdf
16. Pelka, O., Friedrich, C.M.: FHDO biomedical computer science group at medical classification task of ImageCLEF 2015. In: Working Notes of CLEF 2015 (Cross Language Evaluation Forum). CEUR Workshop Proceedings, CEUR-WS.org (September 2015)
17. Rodríguez-Sánchez, A., Fontanella, S., Piater, J., Szedmak, S.: IIS at imageclef 2015: Multi-label classification task. In: Working Notes of CLEF 2015 (Cross Language Evaluation Forum). CEUR Workshop Proceedings, CEUR-WS.org (September 2015)

18. Santosh, K.C.and Xue, Z., Antani, S., Thoma, G.: NLM at imageCLEF2015: Biomedical multipanel figure separation. In: Working Notes of CLEF 2015 (Cross Language Evaluation Forum). CEUR Workshop Proceedings, CEUR-WS.org (September 2015)
19. Tsikrika, T., de Herrera, A.S., Müller, H.: Assessing the scholarly impact of imageclef. In: Cross Language Evaluation Forum (CLEF 2011). Lecture Notes in Computer Science (LNCS), Springer (2011), doi:10.1007/978-3-642-23708-9_12
20. Tsikrika, T., Larsen, B., Müller, H., Endrullis, S., Rahm, E.: The scholarly impact of CLEF (2000–2009). In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pp. 1–12. Springer (2013)
21. 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://ceur-ws.org/Vol-1178/CLEF2012wn-ImageCLEF-VillegasEt2012.pdf
22. Villegas, M., Paredes, R.: Overview of the ImageCLEF 2014 Scalable Concept Image Annotation Task. In: CLEF2014 Working Notes. CEUR Workshop Proceedings, vol. 1180, pp. 308–328. CEUR-WS.org, Sheffield, UK (September 15-18 2014), http://ceur-ws.org/Vol-1180/CLEF2014wn-Image-VillegasEt2014.pdf
23. 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://ceur-ws.org/Vol-1179/CLEF2013wn-ImageCLEF-VillegasEt2013.pdf