Overview of ImageCLEF 2017: Information Extraction from Images

Bogdan Ionescu, Henning Muller, Mauricio Villegas, Helbert Arenas, Giulia Boato, Duc-Tien Dang-Nguyen, Yashin Dicente Cid, Carsten Eickhoff, Alba G. Seco de Herrera, Cathal Gurrin, et al.

Overview of ImageCLEF 2017: Information Extraction from Images

Bogdan Ionescu[1(✉)], Henning Müller[2], Mauricio Villegas[3], Helbert Arenas[4], Giulia Boato[5], Duc-Tien Dang-Nguyen[6], Yashin Dicente Cid[2], Carsten Eickhoff[7], Alba G. Seco de Herrera[8], Cathal Gurrin[6], Bayzidul Islam[9], Vassili Kovalev[10], Vitali Liauchuk[10], Josiane Mothe[4], Luca Piras[11], Michael Riegler[12], and Immanuel Schwall[7]

[1] University Politehnica of Bucharest, Romania
bionescu@alpha.imag.pub.ro
[2] University of Applied Sciences Western Switzerland (HES-SO), Switzerland
[3] SearchInk, Germany
[4] Institut de Recherche en Informatique de Toulouse, France
[5] University of Trento, Italy
[6] Dublin City University, Ireland
[7] ETH Zurich, Switzerland
[8] National Library of Medicine, USA
[9] Technische Universität Darmstadt, Germany
[10] United Institute of Informatics Problems, Belarus
[11] University of Cagliari, Italy
[12] Simula Research Laboratory, Norway

Abstract. This paper presents an overview of the ImageCLEF 2017 evaluation campaign, an event that was organized as part of the CLEF (Conference and Labs of the Evaluation Forum) labs 2017. ImageCLEF is an ongoing initiative (started in 2003) that promotes the evaluation of technologies for annotation, indexing and retrieval for providing information access to collections of images in various usage scenarios and domains. In 2017, the 15th edition of ImageCLEF, three main tasks were proposed and one pilot task: 1) a LifeLog task about searching in LifeLog data, so videos, images and other sources; 2) a caption prediction task that aims at predicting the caption of a figure from the biomedical literature based on the figure alone; 3) a tuberculosis task that aims at detecting the tuberculosis type from CT (Computed Tomography) volumes of the lung and also the drug resistance of the tuberculosis; and 4) a remote sensing pilot task that aims at predicting population density based on satellite images. The strong participation of over 150 research groups registering for the four tasks and 27 groups submitting results shows the interest in this benchmarking campaign despite the fact that all four tasks were new and had to create their own community.

1 Introduction

20 years ago getting access to large visual data sets for research was a problem and open data collections that could be used to compare algorithms of

researchers were rare. Now it is getting easier to access data collections but it is still hard to obtain annotated data with a clear evaluation scenario and strong baselines to compare to. Motivated by this, ImageCLEF has for 15 years been an initiative that aims at evaluating multilingual or language independent annotation and retrieval of images [15,18,5,24]. The main goal of ImageCLEF is to support the advancement of the field of visual media analysis, classification, annotation, indexing and retrieval. It proposes novel challenges and develops the necessary infrastructure for the evaluation of visual systems operating in different contexts and providing reusable resources for benchmarking, which is also linked to initiatives such as Evaluation as a Service (EaaS) [11]. Many research groups have participated over the years in these 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 [22].

There are other evaluation initiatives that have had a close relation with ImageCLEF. LifeCLEF [14] was formerly an ImageCLEF task. However, due to the need to assess technologies for automated identification and understanding of living organisms using data not only restricted to images, but also videos and sound, it was decided to be organised independently from ImageCLEF. Other CLEF labs linked to ImageCLEF, in particular the medical task, are: CLEFeHealth [10] that deals with processing methods and resources to enrich difficult-to-understand eHealth text and the BioASQ [3] tasks from the Question Answering lab that targets biomedical semantic indexing and question answering but is now not a lab anymore. Due to their medical topic, the organisation is coordinated in close collaboration with the medical tasks in ImageCLEF.

This paper presents a general overview of the ImageCLEF 2017 evaluation campaign[1], which as usual was an event organised as part of the CLEF labs[2]. Section 2 presents a general description of the 2017 edition of ImageCLEF, commenting about the overall organisation and participation in the lab. Followed by this are sections dedicated to the four tasks that were organised this year. Section 3 explains all details on the life logging task; Section 4 details the caption prediction task; Section 5 describes the two subtasks for the tuberculosis challenge and the pilot task on remote sensing data is described in Section 6.

For the full details and complete results, the readers should refer to the corresponding task overview papers [6,9,7,2]. The final section of this paper concludes by giving an overall discussion, and pointing towards the challenges ahead and possible new directions for future research.

2 Overview of Tasks and Participation

ImageCLEF 2017 consisted of three main tasks and a pilot task that covered challenges in diverse fields and usage scenarios. In 2016 [25] the tasks were completely different with a handwritten retrieval task, an image annotation task

[1] http://imageclef.org/2017/

[2] http://clef2017.clef-initiative.eu/

and a medical task with several subtasks. In 2017 the tasks completely changed and only the caption prediction was a subtask already attempted in 2016 but for which no participant submitted results in 2016. The 2017 tasks are the following:

- **ImageCLEFlifelog:** aims at developing systems for lifelogging data retrieval and summarization, so for persons automatically logging their life.
- **ImageCLEFcaption:** addresses the problem of bio-medical image caption prediction from large amounts of training data. Captions can either be created as free text or concepts of the image captions could be detected.
- **ImageCLEFtuberculosis:** targets the challenge of determining the tuberculosis (TB) subtypes and drug resistances automatically from the volumetric image information (mainly related to texture) and based on clinical information that is available such as age, gender, etc.
- **ImageCLEFremote (pilot task):** targets the estimation of the population of a geographical area based on low definition but free earth observation images as provided by Copernicus program.

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 a signed End User Agreement (EUA). Table 1 summarizes the participation in ImageCLEF 2017, including the number of registrations and number of signed EUAs, indicated both per task and for the overall lab. The table also shows the number of groups that submitted results (a.k.a. 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 count as a single group participation. The EUA explicitly requires all groups that get access to the data to participate, even though this is not enforced. Unfortunately, the percentage of groups that submit results is often relatively small. Nevertheless, as observed in studies of scholarly impact [22,23], 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.

After a decrease in participation in 2016, the participation increased well in 2017 and this despite the fact that all four tasks did not have a participating community as all tasks were new and had to create the community from scratch. Still, of the 167 groups that registered and 60 that submitted a valid copyright agreement, only 27 submitted results in the end. The percentage is in line with past years with 20% of the registered groups submitting results and about 50% of those that signed the agreement. The following four sections 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.

Table 1: Key figures of participation in ImageCLEF 2017.

Task	Online registrations	Signed EUA	Groups that subm. results	Submitted working notes
Lifelog	66	21	3	3
Caption	100	43	11	11
Tuberculosis	96	40	9	8
Remote	59	20	4	4
Overall	167	60	27	26

3 The Lifelog Task

3.1 Motivation and Task Setup

The availability of a large variety of personal devices, such as smartphones, video cameras as well as wearable devices that allow capturing pictures, videos and audio clips in every moment of our life is creating vast archives of personal data where the totality of an individual's experiences, captured multi-modally through digital sensors are stored permanently as a personal multimedia archive. These unified digital records, commonly referred to as *lifelogs*, gathered increasing attention in recent years within the research community. This happened due to the need for and challenge of building systems that can automatically analyse these huge amounts of data in order to categorize, summarize and also query them to retrieve the information that the user may need.

Despite the increasing number of successful related workshops and panels (e.g., iConf 2016[3] , ACM MM 2016[4]) lifelogging has rarely been the subject of a rigorous comparative benchmarking exercise as, for example, the new lifelog evaluation task at NTCIR-12[5]. The ImageCLEF 2017 LifeLog task [6] aims to bring the attention of lifelogging to a wide audience and to promote research into some of the key challenges of the coming years. The ImageCLEF 2017 LifeLog task aims to be a comparative evaluation of information access and retrieval systems operating over personal lifelog data. The task consists of two sub-tasks, both allow participation independently. These sub-tasks are:

- Lifelog Retrieval Task (LRT);
- Lifelog Summarization Task (LST).

Lifelog retrieval task
The participants had to analyse the lifelog data and according to several specific queries return the correct answers. For example: *Shopping for Wine: Find the moment(s) when I was shopping for wine in the supermarket* or *The Metro: Find*

[3] http://irlld2016.computing.dcu.ie/index.html
[4] http://lta2016.computing.dcu.ie/styled/index.html
[5] http://ntcir-lifelog.computing.dcu.ie/NTCIR12/

Table 2: Statistics of Lifelog Dataset

Number of Lifeloggers	**3**
Size of the Collection (Images)	**88,124** images
Size of the Collection (Locations)	**130** locations
Number of LRT Topics	**36** (16 for devset, 20 for testset)
Number of LsT Topics	**15** (5 for devset, 10 for testset)

the moment(s) when I was riding a metro. The ground truth for this sub-task was created by extending the queries from the NTCIR-12 dataset, which already provides a sufficient ground truth.

Lifelog summarization task
In this sub-task the participants had to analyse all the images and summarize them according to specific requirements. For instance: *Public Transport: Summarize the use of public transport by a user. Taking any form of public transport is considered relevant, such as bus, taxi, train, airplane and boat. The summary should contain all different day-times, means of transport and locations, etc.*

Particular attention had to be paid to the diversification of the selected images with respect to the target scenario. The ground truth for this sub-task was created utilizing crowdsourcing and manual annotations.

3.2 Data Sets Used

The Lifelog dataset consists of data from three lifeloggers for a period of about one month each. The data contains a large collection of wearable camera images (approximately two images per minute), an XML description of the semantic locations (e.g. Starbucks cafe, McDonalds restaurant, home, work) and the physical activities (e.g. walking, transport, cycling), of the lifeloggers at a granularity of one minute. A summary of the data collection is shown in Table 2.

Given the fact that lifelog data is typically visual in nature and in order to reduce the barriers-to-participation, the output of the Caffe CNN-based visual concept detector was included in the test collection as additional meta data.

Topics Aside from the data, the test collection included a set of topics (queries) that were representative of the real-world information needs of lifeloggers. There were 36 and 15 ad-hoc search topics representing the challenge of retrieval for the LRT task and the challenge of summarization for the LST task, respectively.

Evaluation Methodology For the *Lifelog Rerieval Task* evaluation metrics based on NDCG (Normalized Discounted Cumulative Gain) at different depths were used, i.e., $NDCG@N$, where N varies based on the type of the topics, for the recall oriented topics N was larger (> 20), and for the precision oriented topics N was smaller N (5, 10 or 20).

In the *Lifelog Summarization Task* classic metrics were deployed:

– Cluster Recall at X(CR@X) — a metric that assesses how many different clusters from the ground truth are represented among the top X results;
– Precision at X(P@X) — measures the number of relevant photos among the top X results;
– F1-measure at X($F1$@X) — the harmonic mean of the previous two.

Various cut off points were considered, e.g., $X = 5, 10, 20, 30, 40, 50$. Official ranking metrics this year was the **F1-measure@10** or images, which gives equal importance to diversity (via CR@10) and relevance (via P@10).

Participants were also encouraged to undertake the sub-tasks in an interactive or automatic manner. For interactive submissions, a maximum of five minutes of search time was allowed per topic. In particular, the organizers would like to emphasize methods that allowed interaction with real users (via Relevance Feedback (RF), for example), i.e., beside of the best performance, the way of interaction (like number of iterations using RF), or innovation level of the method (for example, new way to interact with real users) has been evaluated.

3.3 Participating Groups and Runs Submitted

We received 18 runs submitted from 3 teams from Singapore, Romania, and a multi-nation team from Ireland, Italy, and Norway. The submitted runs are summarized in Table 3.

3.4 Results

We received approaches from fully automatic to fully manual paradigms, from using a single information provided by the task to using all information as well as extra resources. In Table 4, we report the runs with highest score from each team for both subtasks.

3.5 Lessons Learned and Next Steps

What we learned from the lifelogging task is that multi-modal analysis seems still to be a problem that not many address. Often only one type of data is analysed. For the future it would be important to encourage participants to try out all modalities. This could be achieved by providing pre-extracted features with the data. Apart from that there was a large gap between signed-up teams and submitted runs. We think that this is based on the complexity of the task and the large amount of data that need to be analysed. Supporting participants with pre-extracted features could also help in this case because feature extraction can take much time. Finally, and most importantly, we could show how interesting and challenging lifelog data is and that it holds much research potential, not only in multimedia analysis but also from a system point of view for the performance. For next steps we will enrich the dataset with more data and also look into which pre-extracted features would make sense and what is the best format to share it with our colleagues.

Table 3: Submitted runs for ImageCLEFlifelog 2017 task.

Team	Run	Description
Lifelog Retrieval Subtask.		
Organizers [26]	Baseline	Baseline method, fully automatic.
	Segmentation	Apply segmentation and automatic retrieval based on concepts.
	Fine-tunning	Apply segmentation and fine-tunning. Using all information.
Lifelog Summarization Subtask.		
I2R [17]	Run 1	Parameters learned for maximum F1 score. Using only visual information.
	Run 2	Parameters learned for maximum F1 score. Using visual and metadata information.
	Run 3	Parameters learned for maximum F1 score. Using metadata.
	Run 4	Re-clustering in each iteration; 20% extra clusters. Using visual, metadata and interactive.
	Run 5	No re-clustering. 100% extra clusters. Using visual, metadata and interactive.
	Run 6	Parameters learned for maximum F1 score. Using visual, metadata, and object detection.
	Run 7	Parameters learned for maximum F1 score, w/ and w/o object detection. Using visual, metadata, and object detection.
	Run 8	Parameters learned for maximum F1 score. Using visual information and object detection.
	Run 9	Parameters learned for maximum precision. Using visual and metadata information.
	Run 10	No re-clustering. 20 % extra clusters. Using visual, metadata and interactive.
UPB [8]	Run 1	Textual filtering and word similarity using WordNet and Retina.
Organizers [26]	Baseline	Baseline method, fully automatic.
	Segmentation	Apply segmentation and automatic retrieval and diversification based on concepts.
	Filtering	Apply segmentation, filtering, and automatic diversification. Using all information.
	Fine-tunning	Apply segmentation, fine-tunning, filtering, and automatic diversification. Using all information.
	RF	Relevance feedback. Using all information.

Table 4: ImageCLEFlifelog 2017 results.

Retrieval Subtask.			**Summarization Subtask.**		
Team	**Best Run**	**NDCG**	**Team**	**Best Run**	**F1@10**
Organizers* [26]	Fine-Tuning	0.386	I2R [17]	Run 2	0.497
			UPB [8]	Run 1	0.132
			Organizers* [26]	RF	0.769

*Note: Results from the organizers team are just for reference.

4 The Caption Task

Interpreting and summarizing the insights gained from medical images such as radiography or biopsy samples is a time-consuming task that involves highly trained experts and often represents a bottleneck in clinical diagnosis. Consequently, there is a considerable need for automatic methods that can approximate the mapping from visual information to condensed textual descriptions.

4.1 Task Setup

The ImageCLEF 2017 caption task [9] casts the problem of image understanding as a cross-modality matching scenario in which visual content and textual descriptors need to be aligned and concise textual interpretations of medical images are generated. The task works on the basis of a large-scale collection of figures from open access biomedical journal articles from PubMed Central (PMC)[6]. Each image is accompanied by its original caption and a set of extracted UMLS® (Unified Medical Language System®)[7] Concept Unique Identifiers (CUIs), constituting a natural testbed for this image captioning task.

In 2016, ImageCLEFmed [12] proposed a caption prediction subtask. This edition of the biomedical image captioning task at ImageCLEF comprises two subtasks: (1) Concept Detection and (2) Image Caption Prediction. Figure 1 shows an example biopsy image along with its relevant concepts as well as the reference caption.

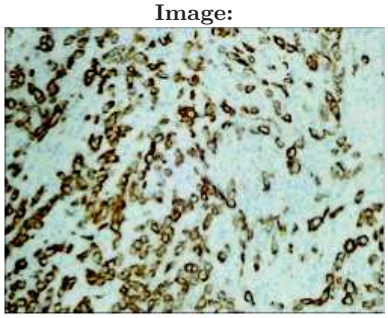
Concept Detection As a first step to automatic image captioning and understanding, participating systems are tasked with identifying the presence of relevant biomedical concepts in medical images. Based on the visual image content, this subtask provides the building blocks for the image understanding step by identifying the individual components from which full captions can be composed.

Caption Prediction On the basis of the concept vocabulary detected in the first subtask as well as the visual information of their interaction in the image, participating systems are tasked with composing coherent natural language captions for the entirety of an image. In this step, rather than the mere coverage of visual concepts, detecting the interplay of visible elements is crucial for recreating the original image caption.

[6] PubMed Central (PMC) is a free full–text archive of biomedical and life sciences journal literature at the U.S. National Institute of Health's National Library of Medicine (NIH/NLM) (see http://www.ncbi.nlm.nih.gov/pmc/).

[7] https://www.nlm.nih.gov/research/umls

Image:



Concept detection:

- C0021067: immunoperoxidase
- C0027651: neoplastia
- C0038128: stain
- C1441616: immunohistochemical

Caption prediction:

Immunohistochemical stain for pancytokeratin, highlighting tumor cells with unstained lymphocytes in the background (*immunoperoxidase stain* x400).

Fig. 1: Example of an image and the information provided in the training set.

4.2 Dataset

The experimental corpus is derived from scholarly biomedical articles of PMC from which we extract figures and their corresponding captions. The collection is comprised of 184,614 image-caption pairs. This overall set is further split into disjunct training (164,614 pairs), validation (10,000 pairs) and test (10,000 pairs) sets. For the concept detection subtask, we used the QuickUMLS library [21] to identify the CUIs mentioned in the caption text.

4.3 Participating Groups and Submitted Runs

We received a total of 71 runs by 11 individual teams. There was a limit of at most 10 runs per team and subtask and the submissions are roughly evenly split between tasks. The vast majority of participating groups relied on some form of neural network architecture, typically combining convolutional and recurrent layers in order to jointly reason about visual and textual information.

4.4 Results

The evaluation of both subtasks is conducted separately. For the concept detection task, we measured the balanced precision and recall trade-off in terms of F_1 scores. Python's scikit-learn (v0.17.1-2) library is used. We compute micro F_1 per image and average across all images. 393 reference captions in the test set do not contain any CUIs. The respective images are excluded from the evaluation.

Caption prediction performance is assessed on the basis of BLEU scores [19] using the Python NLTK (v3.2.2) default implementation. Candidate captions are lower cased, stripped of all punctuation and English stop words. Finally, to increase coverage, we apply Snowball stemming. BLEU scores are computed per reference image, treating each entire caption as a sentence, even though it may contain multiple natural sentences. We report average BLEU scores across all 10,000 test images.

Table 5 gives a detailed performance overview of the concept detection subtask. We differentiate between official runs (O) and those that use external information to train models (E). While the majority of submissions is fully automatic (A), we received a number of runs (M) including some form of manual intervention. Since our entire experimental corpus is in the public domain, the use of external information runs the risk of leaking test images into the training process. For this reason, the task's official ranking concentrates on those teams that relied only on official material and that are therefore directly comparable. The best official results for this task were obtained by Athens University's Information Processing Laboratory.

The results of the caption prediction subtask can be found in Table 6. While there were no manual submissions to this subtask, here, as well, the use of external information gave teams a considerable, yet difficult-to-compare performance advantage and is therefore excluded from the official team ranking. The best official results were obtained by the Chinese Academy of Sciences' Key Laboratory on Intelligent Information Processing (isia). For additional details regarding the participating teams and their approaches, we refer the reader to the task overview paper [9].

4.5 Lessons Learned and Next Steps

There are several observations that need to be taken into account when analyzing the results presented in the previous section. Most notably, as a consequence of the data source (scholarly biomedical journal articles), the collection contains a considerable amount of noise in the form of compound figures with potentially highly heterogeneous content. In future editions of this task, we plan using a more well-defined source of images such as radiology or biopsy samples in order to reduce the amount of variation in the data.

Second, the CUIs extraction employed to generate ground truth labels is a probabilistic process that introduces its own errors. As a consequence, there are a considerable number of training captions that do not contain any CUIs, making such examples difficult to use for concept detection. In the future, plan to rely on more rigorous (manual and thus expensive) filtering to ensure good concept coverage across training, validation and test data.

Finally, the call for contributions did not make any assumptions about the kinds of strategies participants would rely on. As a consequence, we see a broad range of methods being applied. Evaluation of the results shows that some teams employed methods that were at least partially trained on external resources including Pubmed articles. Since such approaches cannot be guaranteed to have

Table 5: Concept detection using official (O) and external (E) resources.

Team	Run	Type	Resources	F_1
NLM	1494012568180	A	E	0.1718
NLM	1494012586539	A	E	0.1648
Aegean AI Lab	1491857120689	A	E	0.1583
Information Processing Laboratory	**1494006128917**	**A**	**O**	**0.1436**
Information Processing Laboratory	1494006074473	A	O	0.1418
Information Processing Laboratory	1494009510297	A	O	0.1417
Information Processing Laboratory	1494006054264	A	O	0.1415
Information Processing Laboratory	1494009412127	A	O	0.1414
Information Processing Laboratory	1494009455073	A	O	0.1394
NLM	1494014122269	A	E	0.1390
Information Processing Laboratory	1494006225031	A	O	0.1365
Information Processing Laboratory	1494006181689	A	O	0.1364
NLM	1494012605475	A	E	0.1228
Information Processing Laboratory	1494006414840	A	O	0.1212
Information Processing Laboratory	1494006360623	A	O	0.1208
AILAB	1493823116836	A	E	0.1208
BMET	1493791786709	A	O	0.0958
BMET	1493791318971	A	O	0.0880
NLM	1494013963830	A	O	0.0880
NLM	1494014008563	A	O	0.0868
BMET	1493698613574	A	O	0.0838
NLM	1494013621939	A	O	0.0811
NLM	1494013664037	A	O	0.0695
Morgan CS	1494060724020	M	O	0.0498
BioinformaticsUA	1493841144834	M	O	0.0488
BioinformaticsUA	1493995613907	M	O	0.0463
mami	1496127572481	M	E	0.0462
Morgan CS	1494049613114	M	O	0.0461
Morgan CS	1494048615677	M	O	0.0434
BioinformaticsUA	1493976564810	M	O	0.0414
Morgan CS	1494048330426	A	O	0.0273
AILAB	1493823633136	A	E	0.0234
AILAB	1493823760708	A	E	0.0215
NLM	1495446212270	A	E	0.0162
MEDGIFT UPB	1493803509469	A	E	0.0028
NLM	1494012725738	A	O	0.0012
mami	1493631868847	M	E	0.0000

Table 6: Caption prediction using official (O) and external (E) resources.

Team	Run	Resources	BLEU
NLM	1494014231230	E	0.5634
NLM	1494081858362	E	0.3317
AILAB	1493825734124	E	0.3211
NLM	1495446212270	E	0.2646
AILAB	1493824027725	E	0.2638
isia	**1493921574200**	**O**	**0.2600**
isia	1493666388885	O	0.2507
isia	1493922473076	O	0.2454
isia	1494002110282	O	0.2386
isia	1493922527122	O	0.2315
NLM	1494038340934	O	0.2247
isia	1493831729114	O	0.2240
isia	1493745561070	O	0.2193
isia	1493715950351	O	0.1953
isia	1493528631975	O	0.1912
AILAB	1493825504037	E	0.1801
isia	1493831517474	O	0.1684
NLM	1494038056289	O	0.1384
NLM	1494037493960	O	0.1131
AILAB	1493824818237	E	0.1107
BMET	1493702564824	O	0.0982
BMET	1493698682901	O	0.0851
BMET	1494020619666	O	0.0826
Biomedical Computer Science Group	1493885614229	E	0.0749
Biomedical Computer Science Group	1493885575289	E	0.0675
BMET	1493701062845	O	0.0656
Biomedical Computer Science Group	1493885210021	E	0.0624
Biomedical Computer Science Group	1493885397459	E	0.0537
Biomedical Computer Science Group	1493885352146	E	0.0527
Biomedical Computer Science Group	1493885286358	E	0.0411
Biomedical Computer Science Group	1493885541193	E	0.0375
Biomedical Computer Science Group	1493885499624	E	0.0365
Biomedical Computer Science Group	1493885708424	E	0.0326
Biomedical Computer Science Group	1493885450000	E	0.0200

respected our division into training, validation and test folds and might subsequently leak test examples into the training process, future editions of the task will carefully describe the categories of submissions based on the resources used.

5 The Tuberculosis Task

About 130 years after the discovery of Mycobacterium tuberculosis, the disease remains a persistent threat and a leading cause of death worldwide. The greatest disaster that can happen to a patient with tuberculosis (TB) is that the organisms become resistant to two or more of the standard drugs. In contrast to drug sensitive (DS) tuberculosis, its multi-drug resistant (MDR) form is more difficult and expensive to treat. Thus, early detection of the drug resistance (DR) status is of great importance for effective treatment. The most commonly used methods of DR detection are either expensive or take too much time (up to several months). Therefore, there is a need for quick and at the same time cheap methods of DR detection. One of the possible approaches for this task is based on Computed Tomography (CT) image analysis. Another challenging task is automatic detection of TB types using CT volumes.

5.1 Task Setup

Two subtasks were then proposed in the ImageCLEF tuberculosis task 2017 [7]:

- Multi-drug resistance detection (MDR subtask);
- Tuberculosis type classification (TBT subtask).

The goal of the MDR subtask is to assess the probability of a TB patient having resistant form of tuberculosis based on the analysis of a chest CT. For the TBT subtask, the goal is to automatically categorize each TB case into one of the following five types: Infiltrative, Focal, Tuberculoma, Miliary, Fibro-cavernous.

5.2 Dataset

For both subtasks 3D CT images were provided with a size of 512 × 512 pixels and number of slices varying from 50 to 400. All CT images were stored in NIFTI file format with .nii.gz file extension (g-zipped .nii files). This file format stores raw voxel intensities in Hounsfield units (HU) as well the corresponding image metadata such as image dimensions, voxel size in physical units, slice thickness, etc. For all patients automatically extracted masks of the lungs were provided.

The dataset for the MDR subtask was composed of 209 MDR and 234 DS patients. The division of the data into training and test sets is shown in Table 7. The TBT task contained 800 patients divided as presented in Table 8. One 2D slice per TB type is shown in Figure 2.

Table 7: Dataset for the MDR subtask.

# Patients	Train	Test
DS	134	101
MDR	96	113
Total patients	230	214

Table 8: Dataset for the TBT subtask.

# Patients	Train	Test
Type 1 (Infiltrative)	140	80
Type 2 (Focal)	120	70
Type 3 (Tuberculoma)	100	60
Type 4 (Miliary)	80	50
Type 5 (Fibro-cavernous)	60	40
Total patients	500	300

5.3 Participating Groups and Submitted Runs

In the first year of the task, 9 groups from 6 countries have submitted at least one run to one of the subtask. There were 8 groups participating in the MDR subtask, and 7 in the TBT task. Each group could submit up to 10 runs. Finally 28 and 23 runs were submitted in the MDR and TBT tasks respectively. 5 groups used a deep-learning approach, two were based on graph models encoding local texture features and one build a co-occurrence of adjacent supervoxels. One group did not explain the algorithm.

5.4 Results

MDR subtask is a 2-class problem. The participants submitted for each patient in the test set the probability of belonging to the MDR group. The Area Under the ROC Curve (AUC) was chosen as the measure to rank results. Accuracy was provided as well. For the TBT subtask, the participants had to submit the tuberculosis category. Since the 5-class problem was not balanced, Cohen's Kappa was used to compare the methods. Again, the accuracy was provided. Tables 9 and 10 show the final results for each run and their rank.

5.5 Lessons Learned and Next Steps

The results underline the difficulty of both tasks. In the case of the MDR task all participants were close to an AUC of 0.50 that is the performance of a random classifier. When considering the accuracy the results are sometimes worse. The random accuracy for this subtask is 0.5280 and the best participant reached an accuracy of 0.5681. In the TBT subtask the results are more promising. 6 runs achieved a Cohen's Kappa of better than 0.21, threshold to consider a fair agreement between classifications. The random accuracy in this case would be 0.2667 and most of the participants were above this value.

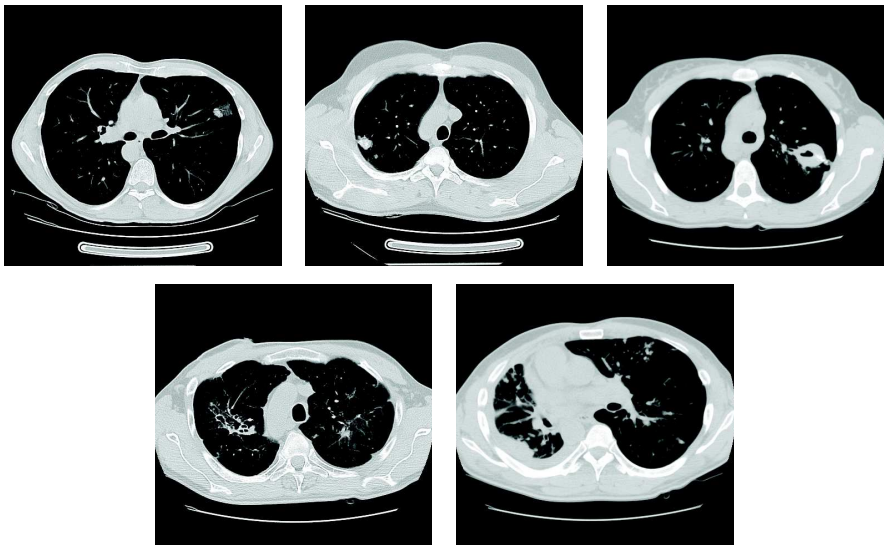


Fig. 2: Examples of the TB types. First row, from left to right: Infiltrative, Focal, and Tuberculoma types. Second row: Miliary, and Fibro-cavernous types.

The analysis of the results and the different nature of the methods suggest that the training data did not fully represent the test cases, being fairly small for the diversity of the cases. In the MDR subtask the set of DS patients was composed of patients that may have presented resistance to some drugs, but no all. With more training cases, the groups can be better defined. In a future edition of this task we expect to add the current test set as training and provide new patients for the test set.

Table 9: Results for the MDR subtask.

Group Name	Run	AUC	ACC	Rank
MedGIFT	MDR_Top1_correct.csv	0.5825	0.5164	1
MedGIFT	MDR_submitted_topBest3_correct.csv	0.5727	0.4648	2
MedGIFT	MDR_submitted_topBest5_correct.csv	0.5624	0.4836	3
SGEast	MDR_LSTM_6_probs.txt	0.5620	0.5493	4
SGEast	MDR_resnet_full.txt	0.5591	0.5493	5
SGEast	MDR_BiLSTM_25_wcrop_probs.txt	0.5501	0.5399	6
UIIP	MDR_supervoxels_run_1.txt	0.5415	0.4930	7
SGEast	MDR_LSTM_18_wcrop_probs.txt	0.5404	0.5540	8
SGEast	MDR_LSTM_21wcrop_probs.txt	0.5360	0.5070	9
MedGIFT	MDR_Top2_correct.csv	0.5337	0.4883	10
HHU DBS	MDR_basecnndo_212.csv	0.5297	0.5681	11
SGEast	MDR_LSTM_25_wcrop_probs.txt	0.5297	0.5211	12
BatmanLab	MDR_submitted_top5.csv	0.5241	0.5164	13
HHU DBS	MDR_basecnndo_113.csv	0.5237	0.5540	14
MEDGIFT UPB	MDR_TST_RUN_1.txt	0.5184	0.5352	15
BatmanLab	MDR_submitted_top4_0.656522.csv	0.5130	0.5024	16
MedGIFT	MDR_Top3_correct.csv	0.5112	0.4413	17
HHU DBS	MDR_basecnndo_132.csv	0.5054	0.5305	18
HHU DBS	MDR_basecnndo_182.csv	0.5042	0.5211	19
HHU DBS	MDR_basecnndo_116.csv	0.5001	0.4930	20
HHU DBS	MDR_basecnndo_142.csv	0.4995	0.5211	21
HHU DBS	MDR_basecnndo_120.csv	0.4935	0.4977	22
SGEast	MDR_resnet_partial.txt	0.4915	0.4930	23
BatmanLab	MDR-submitted_top1.csv	0.4899	0.4789	24
BatmanLab	MDR_SuperVx_Hist_FHOG_rf_0.648419.csv	0.4899	0.4789	25
Aegean Tubercoliosis	MDR_DETECTION_EXPORT2.csv	0.4833	0.4648	26
BatmanLab	MDR_SuperVx_FHOG_rf_0.637994.csv	0.4601	0.4554	27
BioinformaticsUA	MDR_run1.txt	0.4596	0.4648	28

6 The Remote (Population Estimation) Task

6.1 Motivation and Task Setup

Before engaging any rescue operation or humanitarian action, NGOs (Non-Governmental Organizations) need to estimate the local population as accurately as possible. Population estimation is fundamental to provide any service for a particular region. While good estimates exists in many parts of the world through accurate census data, this is usually not the case in developing countries.

This pilot task, introduced in 2017, aims at investigating the use of satellite data as a cheaper and quicker process. The task uses Copernicus Sentinel-2 images with resolution between 10 to 60 meters.

6.2 Data Sets Used

In this pilot task, participants had to estimate the population for different areas in two regions. To achieve this goal, organizers provided a set of satellite images

Table 10: Results for the TBT subtask.

Group Name	Run	Kappa	ACC	Rank
SGEast	TBT_resnet_full.txt	0.2438	0.4033	1
SGEast	TBT_LSTM_17_wcrop.txt	0.2374	0.3900	2
MEDGIFT UPB	TBT_T_GNet.txt	0.2329	0.3867	3
SGEast	TBT_LSTM_13_wcrop.txt	0.2291	0.3833	4
Image Processing	TBT-testSet-label-Apr26-XGao-1.txt	0.2187	0.4067	5
SGEast	TBT_LSTM_46_wcrop.txt	0.2174	0.3900	6
UIIP	TBT_iiggad_PCA_RF_run_1.txt	0.1956	0.3900	7
MEDGIFT UPB	TBT_...._GoogleNet_10crops_at_different_scales_.txt	0.1900	0.3733	8
SGEast	TBT_resnet_partial.txt	0.1729	0.3567	9
MedGIFT	TBT_Top1_correct.csv	0.1623	0.3600	10
SGEast	TBT_LSTM_25_wcrop.txt	0.1548	0.3400	11
MedGIFT	TBT_submitted_topBest3_correct.csv	0.1548	0.3500	12
BatmanLab	TBT_SuperVx_Hist_FHOG_lr_0.414000.csv	0.1533	0.3433	13
SGEast	TBT_LSTM_37_wcrop.txt	0.1431	0.3333	14
MedGIFT	TBT_submitted_topBest5_correct.csv	0.1410	0.3367	15
MedGIFT	TBT_Top4_correct.csv	0.1352	0.3300	16
MedGIFT	TBT_Top2_correct.csv	0.1235	0.3200	17
BatmanLab	TBT_submitted_bootstrap.csv	0.1057	0.3033	18
BatmanLab	TBT_submitted_top3_0.490000.csv	0.1057	0.3033	19
BatmanLab	TBT_SuperVx_Hist_FHOG_Reisz_lr_0.426000.csv	0.0478	0.2567	20
BatmanLab	TBT_submitted_top2_0.430000.csv	0.0437	0.2533	21
BioinformaticsUA	TBT_run0.txt	0.0222	0.2400	22
BioinformaticsUA	TBT_run1.txt	0.0093	0.1233	23

(Copernicus Sentinel 2)[8]. The boundaries of the areas of interest were provided as shape files. The clipped satellite images were provided as well as the meta data of the original images (before clipping). The data set consists of topographic and geographic information as follows:

- ESRI (Environmental Systems Research Institute) shape files: there is a single shape file by region and the projected shape file of the region has the attributes to represent the various areas the region is composed of.
- Sentinel-2 satellite images: The remote sensing imagery are from the Sentinel-2 platform. The imagery is multi spectral, cloud-free satellite imagery downloaded from Sentinel Data Hub[9]. The images have been clipped to match the bounding box of the areas of interest. The bands for images have different spatial resolutions: 10 meters for bands B2 (490nm), B3 (560nm) B4 (665 nm) and B8 (84nm); 20 meters for bands B5 (705nm), B6 (749nm) B7 (783nm), B8a (865nm) B11 (1610nm) and B12 (2190nm). For the analysis, participants were encouraged to use Red, Green and Blue bands or in some cases near infrared bands that are 10 meters in resolution.

[8] The dataset is available on Zenodo with the DOI 10.5281/zenodo.804602 or on demand

[9] https://scihub.copernicus.eu/dhus/#/home

– Meta-data associated to the images: Information regarding the original images is provided in XML files. These files contain information like capture time/date, sensor mode, orbit number, the id of quality files, etc. Further information regarding the Sentinel-2 products, as well as file structure can be found in the Sentinel 2 User handbook[10].

However, participants were allowed to use any other resource they think might help to reach the highest accuracy.

There were 83 areas of interest in the city of Lusaka and 17 in west Uganda for which the population has to be estimated. For 90 of these 100 areas, ground truth provided by NGOs is available, so evaluation considered these areas.

Runs from participants are evaluated against ground truth. For the city of Lusaka, the ground truth comes with a categorical evaluation measure of the population estimation, Good (23 over the 83 areas), Acceptable (37), Doubts (9), High doubts (6) and Unknown (8). For West Uganda, the ground truth corresponds to estimations that are based on a combination of Volunteered Geographic information (VGI) working on BING imagery (2012) with additional ground work. Both have been provided by NGOs.

In our evaluation we use three metrics: 1) Sum of differences, which corresponds to the sum of the absolute value of the difference between ground truth and the estimated population over the areas, 2) Root Mean Square Error (RMSE), which is computed as the square root of the average of squared errors [4], 3) Pearson correlation, and 4) AvgRelDelta, which is the average of the relative deltas as calculated in 1) relative to the population.

The four measures that are detailed in [2], aim at comparing the estimated value against the ground truth. The challenge comprise two areas geographically separated, Uganda, and Zambia. Because of this fact, it was decided to evaluate both areas separately. We evaluate the results on two variables: 1) Population counts, and 2) Dwelling counts. Then each run submitted by the participants has 12 possible metrics. However, not all the participants submitted results for both variables. All the submissions provided estimation for the population, while only two of them provided estimates for both population and dwelling counts.

6.3 Participating Groups and Runs Submitted

Although the pilot task was open to anyone, participants came from local hackathon -like events that were organized within FabSpace 2.0 project (`https://fabspace.eu/`); see [1] for details. There are four groups participating with their contribution being summarized in Table 11.

6.4 Results

As can be seen in Table 12, the sum of deltas in the prediction over the 90 areas is in the same range for the 3 participants. The correlation is not very high leaving room for improvement. More details are provided in the task overview.

Table 11: Participants of the ImageCLEF remote task.

Run	Approach
Darmstadt [13]	Supervised (Maximum Likelihood) by false colour composite and NIR band and unsupervised (K-Means Cluster Analysis by NIR band
Grapes [16]	Supervised classification on Sentinel 2 images coupled with statistical forecasting on historical census data
FABSPACE PL	Pre- processing Sentinel-1 data, creating mask with buildings, mean- shift segmentation process
AndreaDavid [20]	Convolutional Neural Network with Sentinel 2 and open data

Table 12: Results on the estimation of the population over the 90 areas from the two regions. Detailed results can be found in [2]. Bold font highlights the best result while the italic font highlights the second best.

Participants	Country	Sum Delta	RMSE	Pearson	AvgRelDelta
Darmstadt [13]	Germany	1,493,152	27,495	0.22	*97.89*
Grapes [16]	Greece	1,486,913	34,290	*0.33*	177.55
FABSPACE PL	Poland	1,558,639	31,799	**0.37**	172.84
AndreaDavid [20]	Italy	1,484,088	27,462	0.21	**87.57**

Figure 3 shows an overview of the results submitted by the participating teams. The maps show the prediction errors divided by the ground truth population for each operational zone ($delta = (d_t - s_t)/d_t.100$).

Operational zones where the models severely overestimated the population (the models suggest a higher population) are shown in red or orange. We consider results severely overestimated when the population estimation is over 50% of the actual population. Areas in which the estimation is +/− 50% are depicted in green, while areas in which the models severely underestimated the population are depicted in blue. In this paper, a population estimation would be considered severely underestimated if it is lower than 50% of the actual population.

We can see that there are areas overestimated by all the models: the Industrial area (West), Ngwerere (North) and Libala (South). In the case of the industrial areas it seems that the proposed algorithms confused industrial buildings with residential areas. In the case of Ngwere, and Libala, the residential areas have low density, which was incorrectly evaluated by the algorithms.

On the other hand, we can see that there are other areas that are underestimated by all the models: George, Lilanda, Desai, (at the West of the city), Chelston at the East, Chawama and Kuoboka at the South. Most of the models underestimated the population in Makeni, except for the Polish team. This team also differentiated from the rest in an area comprised by Ngombe, Chamba

[10] https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook

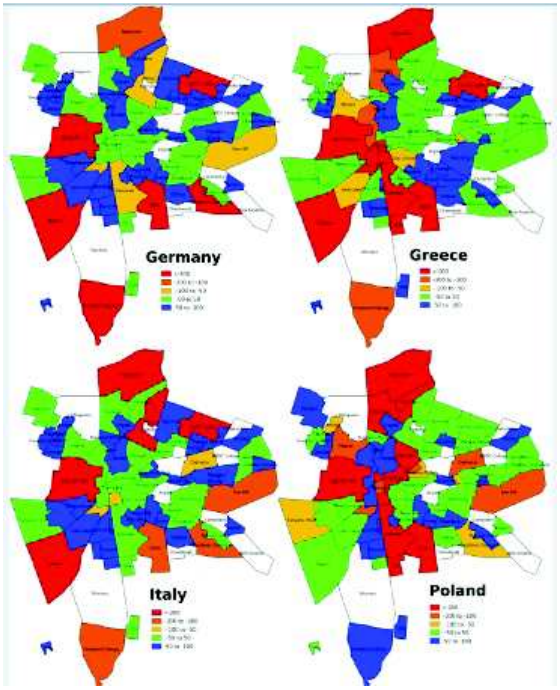


Fig. 3: Overview of the results submitted by the participating teams for the operational zones in Lusaka- Zambia.

Valley, Kamanga and Kaunda square (North East of the city), providing good estimates with the exception of Chudleigh, which was overestimated by all the teams, except for the Italian team.

In general, all the teams obtained best results in an area near the center of the city, an area roughly defined by the Operational zones, Civic Centre, Rhodes Park and in most cases Northmead (except for the Polish team that did not provide a good result for this zone).

6.5 Lessons Learned and Next Steps

One objective of this pilot task was to evaluate the accuracy of population estimation based on low definition images. The motivation is mainly the availability of such images free of charge, for any place and with a high refresh rate of 5 days) thanks to the European Copernicus program. Participants encountered difficulties mainly linked to the nature of the images. The results show that the accuracy needs to be improved to be fully usable in real applications. The time allowed to solve this task was certainly not sufficient and requires good knowledge of multispectral image analysis, which not all participants had. Thanks to the pilot task we now have several ways to improve the estimation.

7 Conclusions

This paper presents a general overview of the activities and outcomes of the ImageCLEF 2017 evaluation campaign. Four tasks were organised covering challenges in: lifelog retrieval, caption prediction, tuberculosis type and drug resistance detection and of remote sensing.

The participation increased compared to previous years with over 160 registered participants and in the end 27 groups submitting results and a large number of runs. This is remarkable as all four tasks were new and had to create a new community. Whereas several of the participants had participated in the past there was also a large number of groups totally new to ImageCLEF and also collaborations of research groups in several tasks.

Deep Neural Networks were applied for basically all tasks and often led to very good results but this was not true for all tasks as graph-based approaches led to best results in the MDR tuberculosis task. The caption prediction task created a large variety of approaches including using content-based image retrieval to find the visually most similar figures for predicting a caption of an image in addition to the visual content itself. The task also showed that it is important to group the submission based on the resources used as external resources can lead to much better results and a comparison of the techniques needs to be based on the same types of resources used.

ImageCLEF 2017 again brought together an interesting mix of tasks and approaches and we are looking forward to the discussions at the workshop.

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC). It is also partly supported European Union's Horizon 2020 Research and Innovation programme under the Grant Agreement n°693210 (FabSpace 2.0).

References

1. Arenas, H., Baker, A., Bialczak, A., Bargiel, D., Becker, M., Gaildrat, V., Carbone, F., Heising, S., Islam, M.B., Lattes, P., Marantos, C., Menou, C., Mothe, J., Nzeh Ngong, A., Paraskevas, I.S., Penalver, M., Sciana, P., Soudris, D.: FabSpaces at Population Estimation (Remote) Task - ImageCLEF at the CLEF 2017 Labs. CLEF working notes, CEUR (September 11-14 2017)
2. Arenas, H., Islam, B., Mothe, J.: Overview of the ImageCLEF 2017 Population Estimation Task. In: CLEF 2017 Labs Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)
3. Balikas, G., Krithara, A., Partalas, I., Paliouras, G.: BioASQ: A challenge on large–scale biomedical semantic indexing and question answering. In: Multimodal Retrieval in the Medical Domain (MRMD) 2015. Lecture Notes in Computer Science, Springer (2015)

4. Barnston, A.G.: Correspondence among the correlation, rmse, and heidke forecast verification measures; refinement of the heidke score. Weather and Forecasting 7(4), 699–709 (1992)
5. Clough, P., Müller, H., Sanderson, M.: The CLEF 2004 cross–language image retrieval track. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign. Lecture Notes in Computer Science (LNCS), vol. 3491, pp. 597–613. Springer, Bath, UK (2005)
6. Dang-Nguyen, D.T., Piras, L., Riegler, M., Boato, G., Zhou, L., Gurrin, C.: Overview of ImageCLEFlifelog 2017: Lifelog Retrieval and Summarization. In: CLEF 2017 Labs Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)
7. Dicente Cid, Y., Kalinovsky, A., Liauchuk, V., Kovalev, V., , Müller, H.: Overview of ImageCLEFtuberculosis 2017 - predicting tuberculosis type and drug resistances. In: CLEF 2017 Labs Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)
8. Dogariu, M., Ionescu, B.: A Textual Filtering of HOG-based Hierarchical Clustering of Lifelog Data. CLEF working notes, CEUR (September 11-14 2017)
9. Eickhoff, C., Schwall, I., García Seco de Herrera, A., Müller, H.: Overview of ImageCLEFcaption 2017 - image caption prediction and concept detection for biomedical images. In: CLEF 2017 Labs Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)
10. 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)
11. Hanbury, A., Müller, H., Balog, K., Brodt, T., Cormack, G.V., Eggel, I., Gollub, T., Hopfgartner, F., Kalpathy-Cramer, J., Kando, N., Krithara, A., Lin, J., Mercer, S., Potthast, M.: Evaluation–as–a–service: Overview and outlook. ArXiv 1512.07454 (2015)
12. García Seco de Herrera, A., Schaer, R., Bromuri, S., Müller, H.: Overview of the ImageCLEF 2016 medical task. In: Working Notes of CLEF 2016 (Cross Language Evaluation Forum) (September 2016)
13. Islam, M.B., Becker, M., Bargiel, D., Ahmed, K.R., Duzak, P., Emana, N.G.: Sentinel-2 satellite imagery based population estimation strategies at FabSpace 2.0 Lab Darmstadt (September 11-14 2017)
14. Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.P., Lombardo, J.C., Planqué, R., Palazzo, S., Müller, H.: Lifeclef 2017 lab overview: multimedia species identification challenges. In: Proceedings of CLEF 2017 (2017)
15. Kalpathy-Cramer, J., García Seco de Herrera, A., Demner-Fushman, D., Antani, S., Bedrick, S., Müller, H.: Evaluating performance of biomedical image retrieval systems: Overview of the medical image retrieval task at ImageCLEF 2004–2014. Computerized Medical Imaging and Graphics 39(0), 55 – 61 (2015)
16. Koutsouri, K., Skepetari, I., Anastasakis, K., Lappas, S.: Population estimation using satellite imagery. In: CLEF working notes, CEUR. CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)
17. Molino, A.G.D., Mandal, B., Lin, J., Lim, J.H., Subbaraju, V., Chandrasekhar, V.: VC-I2R@ImageCLEF2017: Ensemble of Deep Learned Features for Lifelog Video Summarization. CLEF working notes, CEUR (September 11-14 2017)

18. Müller, H., Clough, P., Deselaers, T., Caputo, B. (eds.): ImageCLEF – Experimental Evaluation in Visual Information Retrieval, The Springer International Series On Information Retrieval, vol. 32. Springer, Berlin Heidelberg (2010)
19. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
20. Pomente, A., Aleandri, D.: Convolutional expectation maximization for population estimation. CLEF working notes, CEUR (September 11-14 2017)
21. Soldaini, L., Goharian, N.: Quickumls: a fast, unsupervised approach for medical concept extraction. In: MedIR Workshop, SIGIR (2016)
22. Tsikrika, T., García Seco de Herrera, A., Müller, H.: Assessing the scholarly impact of ImageCLEF. In: CLEF 2011. pp. 95–106. Springer Lecture Notes in Computer Science (LNCS) (sep 2011)
23. 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)
24. Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., García Seco de Herrera, A., Bromuri, S., Amin, M.A., Kazi Mohammed, M., Acar, B., Uskudarli, S., Marvasti, N.B., Aldana, J.F., Roldán García, M.d.M.: General overview of ImageCLEF at the CLEF 2015 labs. In: Working Notes of CLEF 2015. Lecture Notes in Computer Science, Springer International Publishing (2015)
25. Villegas, M., Müller, H., Garcia Seco de Herrera, A., Schaer, R., Bromuri, S., Gilbert, A., Piras, L., Wang, J., Yan, F., Ramisa, A., Dellandrea, A., Gaizauskas, R., Mikolajczyk, K., Puigcerver, J., Toselli, A.H., Sanchez, J.A., Vidal, E.: General overview of ImageCLEF at the CLEF 2016 labs. In: CLEF 2016 Proceedings. Lecture Notes in Computer Science, Springer, Evora. Portugal (September 2016)
26. Zhou, L., Piras, L., Rieger, M., Boato, G., Dang-Nguyen, D.T., Gurrin, C.: Organizer Team at ImageCLEFlifelog 2017: Baseline Approaches for Lifelog Retrieval and Summarization. CLEF working notes, CEUR (September 11-14 2017)