Image Retrieval Evaluation In Specific Domains

Luca Piras, Barbara Caputo, Duc-Tien Dang-Nguyen, Michael Riegler and Pål Halvorsen

Abstract Image retrieval was, and still is, a hot topic in research . It comes with many challenges that changed over the years with the emergence of more advanced methods for analysis and enormous growth of images created, shared and consumed. This chapter gives an overview of domain-specific image retrieval evaluation approaches, which were part of the ImageCLEF evaluation campaign . Specifically, the robot vision , photo retrieval, scalable image annotation and lifelogging tasks are presented . The ImageCLEF medical activity is described in a separate chapter in this volume. Some of the the presented tasks have been available for several years, whereas others are quite new (like lifelogging). This mix of new and old topics has been chosen to give the reader an idea about the development and trends within image retrieval. For each of the tasks, the datasets, participants, techniques used and lessons learned are presented and discussed leading to a comprehensive summary.

Luca Piras
University of Cagliari, Cagliari, Italy, e-mail: luca.piras@diee.unica.it
Pluribus One, Cagliari, Italy, e-mail: luca.piras@pluribus-one.it

Barbara Caputo
University of Rome La Sapienza, Rome, Italy e-mail: caputo@diag.uniroma1.it

Duc-Tien Dang-Nguyen
University of Bergen, Bergen, Norway e-mail: ductien.dangnguyen@uib.no
Dublin City University, Dublin, Ireland

Michael Riegler
Simula Metropolitan Center for Digital Engineering and University of Oslo, Oslo, Norway e-mail: michael@simula.no

Pål Halvorsen
Simula Metropolitan Center for Digital Engineering, Simula Research Laboratory and University of Oslo, Oslo, Norway e-mail: paalh@simula.no

1 Introduction

In todays modern society, billions of people produce, upload and share digital images using devices like mobile phones or tablet computers. Autonomous intelligent machines like robots, drones and self-driving cars are equipped with RGB-D cameras, continuously capturing visual data to provide on-the-fly information about where the agent is, to subsequently inform its actions. Thus, images and more generally visual data have become a more important and natural way of communication ("a picture says more than thousand words"). This development is leading to large amounts of image data. Flickr, which started in 2004, reported until December 2017, a total number of around 6.5 billion uploaded photos, and Facebook reports around 300 million uploaded images per day. As one can see, there is an immanent need for methods supporting users with their image collections, and artificial machines with the understanding of their visual data. Retrieving images with a particular content, matching a particular query or for a particular purpose from large collections is challenging and has been a focus of research for much time. Likewise, recognizing objects, landmarks and scenes regardless of the imaging conditions are among the holy grails of computer vision since its infancy.

The ImageCLEF initiative and its community became aware of these emerging challenges at an early stage, and Image annotation and retrieval tasks have been part of ImageCLEF since 2003 (Clough and Sanderson, 2004), while the Robot vision challenge was added in 2010 (Pronobis et al, 2010b). In this chapter, we provide an overview of the tasks over the years and how they developed. This gives a unique insight into how image retrieval and robot visual scene understanding changed over the years and which challenges emerged on the road.

In the early years, the focus was on retrieving relevant images from a web-collection given (multi-lingual) queries (Clough et al, 2005, 2006). From 2006 onwards, annotation tasks were also held, initially aimed at object detection (Clough et al, 2007; Deselaers et al, 2008), and more recently, also covering semantic concepts (Deselaers and Hanbury, 2009; Nowak and Dunker, 2010; Nowak and Huiskes, 2010; Nowak et al, 2011; Thomee and Popescu, 2012; Villegas and Paredes, 2012; Villegas et al, 2013; Villegas and Paredes, 2014), photo retrieval (Grubinger et al, 2008; Arni et al, 2009; Lestari Paramita et al, 2010; Zellhöfer, 2012, 2013), and robot vision (Pronobis et al, 2010b,a; Martínez-Gómez et al, 2012, 2013, 2014). In the last editions (Gilbert et al, 2015, 2016), the image annotation task was expanded to concept localization and also natural language sentential description of images. In recent years, there has been an increased interest in research combining text and vision. Therefore, in 2017, there has been a slight change in the focus of the retrieval goal. The task aimed at further stimulating and encouraging multi-modal research that uses text and visual data, and natural language processing for image retrieval and summarization (Dang-Nguyen et al, 2017b).

Although the tasks presented in this chapter are very diverse, spanning a large range of communities traditionally separated, it is still possible to identify a few crucial, shared lessons learned across the tasks over the years:

- multi-modal analysis improves performance compared to single-modal, but multi-modal is rarely exploited by researchers,
- methods change over time (for example switch to deep neural networks), but methods alone cannot solve the challenge entirely, and
- the larger the datasets gets, the harder it gets for participants to process them, whereas on the other side, too small datasets are also not interesting since they are not very applicable for deep learning.

All in all, image retrieval and visual place understanding for robotics applications still holds a lot of open research challenges that go far beyond simple classification such as semantics, object and landmarks detection and recognition, intent and personalized archives.

In the following sections, a detailed overview of selected tasks in the recent years is given. For each task, the data, participants, methods used and lessons learned are presented and discussed.

2 Tasks, Data and Participation

2.1 Overview of the Robot Vision Task

The robot vision challenge was first organized in 2009, as part of the ImageCLEF lab (Pronobis et al, 2010b). Since then, the challenge has been organized another four consecutive times (Pronobis et al, 2010b,a; Martínez-Gómez et al, 2012, 2013, 2014). The first challenge focused on image-based place categorization, namely how to determine from a single image which room the robot is in. In its initial editions, the task focused exclusively on the use of RGB images for place categorization, and the participants were asked to process each image individually. Over the five years of the competition, the challenge grew in complexity so to include multi-modal data, and also in terms of the specific classification tasks required of the participants .

A very strong drive behind the organization of the Robot Vision Challenge was the need to provide the robot vision community with a benchmark where to measure quantitatively progress in semantic localization over the years. Indeed, performing repeatable experiments which produce quantitative, comparable results is a major challenge in robotics for many reasons. To begin with, running experiments often requires expensive hardware. Historically, such hardware has been almost always custom built and standardized, and complete robot platforms started to emerge only recently. Moreover, executing experiments involving real robots is often very time consuming and can be a major engineering challenge. As a result, a large chunk of robotics research has been evaluated in simulation or on a very limited scale. By offering standardized benchmarks and publicly available databases, the Robot Vision Challenge has provided a tool allowing for fair comparisons, simplification of the experimental process, and as a result, a boost for progress in the field of robot vision.

Table 1 Task evolution in the Robot Vision Challenge.

		1st Edition	2nd Edition	3rd Edition	4th Edition	5th Edition
Sources						
	Monocular Images	X	-	-	X	X
	Stereo Images	-	X	X	-	-
	Depth Images	-	-	-	X	-
	Point Clouds	-	-	-	-	X
	Semantic Annotations	X	X	X	X	X
	Pose Annotations	X	-	-	-	-
Objectives						
	Two Tasks	X	X	X	X	-
	Unknown Classes	-	X	X	-	-
	Kidnapping	-	-	-	X	-
	Object Detection	-	-	-	-	X

2.1.1 Datasets

The Robot Vision Challenge was initially conceived as a visual place recognition competition, and the vision component has remained very strong in all its editions. Still, over the years, other additional tasks have been included. Table 1 illustrates these changes.

Accordingly, several datasets have been created over the years. The first dataset used in the challenge was the KTH-IDOL2 database (Luo et al, 2007). It was acquired using a mobile robot platform in the indoor environment of The Computer Vision and Active Perception laboratory (CVAP) at The Royal Institute of Technology (KTH) in Stockholm, Sweden. Each training image was annotated with the topological location of the robot and its pose $< x; y; \theta >$. Although the pose information was provided in the training data, participants were discouraged from using it in their final submission. The two editions of the competition that took place in 2010 were based on COLD-Stockholm, an extension of *COsy Localization Database (COLD)* (Pronobis and Caputo, 2009). This dataset was generated using a pair of high-quality cameras for stereo vision inside the same environment, as for the KTH-IDOL2 dataset. The fourth edition of the challenge used images from the unreleased VIDA dataset (Martínez-Gómez et al, 2013). This dataset includes perspective and range images acquired with a Kinect camera at the Idiap Research Institute in Martigny, Switzerland. Depth information was provided in the form of depth images, with color codes used to represent different distances. Finally, the fifth edition of the competition used images from the unreleased dataset ViDRiLO: The Visual and Depth Robot Indoor Localization with Objects information Dataset. This dataset includes images of the environment and point cloud files (in PCD format) (Martínez-Gómez et al, 2014). Table 2 summarizes the number of classes, as well as the number of training, validation and test images in each edition of the

Table 2 Number of classes and training, validation and test set size.

Task edition	Number of classes	Training images	Validation images	Test images
1st	5	2899	2789	1690
2nd	9	12684	4783	5102
3rd	10	4782	2069	2741
4th	9	7112	0	6468
5th	10	5263	1869	3515

competition. It is worth underlining that the second and third editions included an unknown class not imaged in the training/validation sequences.

2.1.2 Evaluation Measures

The Robot Vision Challenge has always been focused on two main tasks, focused on visual place recognition. In the first one (mandatory task), participants have to provide information about the location of the robot for each test image from the data sequence perceived by the robot separately, i.e. without making any use of the label assigned to the image acquired at time t to make any prediction about the label to be assigned to the image at the time $t+1$. In the second, optional task, instead the temporal continuity of the sequence can be used to improve the final classification of all images . The fifth edition of the challenge also introduced an object recognition task . Note that visual place recognition and object recognition can be considered as two subproblems of semantic localization, where each location is described in terms of its semantic contents.

As evaluation measure, all the editions used a score that computed the performance of the participant submission. This score was always based on positive values for test images correctly classified and negative values for misclassified ones. We also allowed the possibility to not classify test images, resulting in a non-effect on the score. The maximum reachable scores for the mandatory task of each edition were 1,690, 5,102, 2,741, 2,445, and 7,030, respectively. Regarding the optional task, the maximum scores were 1,690, 5,102, 2,741, and 4,079, respectively, for the first to fourth editions. The fifth edition of the task had no optional task.

2.1.3 Participants and Submissions

For all editions of the Robot Vision Challenge, a large number of groups registered, but only a small percentage of them actually participated in the competition and submitted results (see Table 3). We see that the number of registered groups grew considerably over the years, starting with 19 groups at the first edition, having a peak of 71 registered groups in third edition, and somehow reaching a plateau of roughly 40 groups registered in last two edition. In contrast, the number of groups actually participating in the challenges has been more stable over the years, with

an average of 7 groups submitting their runs for the actual challenge every year. The submitted working notes were even less, with the highest number of working notes submitted in the first edition of the task (5 working notes submitted), and the minimum number reached for the last edition (2 working notes submitted).

Table 3 Participation to the Robot Vision Challenge over the years.

Participation	1st Edition	2nd Edition	3rd Edition	4th Edition	5th Edition
Registered Groups	19	28	71	43	39
Participant Groups	7	8	7	8	6
Working Notes Submitted	5	3	3	4	2

2.1.4 Techniques used

Nineteen different groups registered in the first edition of the Robot Vision Challenge (Pronobis et al, 2010b), organized in 2009. For the mandatory task, a wide range of techniques were proposed for the image representation and classification steps . The best result, 793 points out of 1,690, was obtained by the Idiap group using a multi-cue discriminative approach (Xing and Pronobis, 2009). The visual cues considered by this group included global as well as local descriptors, and then an *Support Vector Machines (SVM)* was trained for each visual cue and a high-level cue integration scheme *Discriminative Accumulation Scheme (DAS)* (Nilsback and Caputo, 2004) was used to combine the scores provided by the different SVMs.

For the optional task, the best result, 916.5 points, was obtained by the SIMD group (Martínez-Gómez et al, 2009) using a particle filter approach to estimate the position of the robot given the previous position.

The 2010@ICPR edition (Pronobis et al, 2010a) had a participation similar to the first edition. Amongst the several proposals for the mandatory task, the approach adopted by the CVG group (Fraundorfer et al, 2010) stood out for its full usage of the stereo images to reconstruct the 3D geometry of the rooms, a choice that allowed them to achieve the best score of 3,824 points out of 5,102. The optional task was once again won by the SIMD group (Martínez-Gómez et al, 2010), where they this time computed similarities among local features between test frames and a set of training candidate frames, which was selected by means of clustering techniques . In addition, a sort of temporal smoothing using prior assigned labels was used to classify very uncertain test frames.

The third edition of the Robot Vision challenge (Agosti et al, 2010) required the competing algorithms to show higher generalization capabilities compared to the previous editions. For the second time in a row, the mandatory task was won by the CVG group, with an approach combining a weighted k-NN search using global features, with a geometric verification step (Saurer et al, 2010). This approach obtained a score of 677 points out of 2,741. For the optional task, the approach proposed by

the Idiap group (Fornoni et al, 2010) was to be the most effective. The proposed multi-cue system combined up to three different visual descriptors in a discriminative multiple-kernel SVM. A door detector was implemented for discovering the transition from one room to another, while a stability estimation algorithm was used to evaluate the stability of the classification process.

As mentioned above, the 2012 edition (Martínez-Gómez et al, 2013) of the task introduced range images obtained with a Microsoft Kinect sensor. The organizers proposed a baseline method for both the feature extraction and the classification steps . The group from the Universidad Tecnologica Nacional, Cordoba, Argentina (CIII UTN FRC) (Sánchez-Oro et al, 2013) was the winner for both the mandatory and optional tasks with a score of 2,071 and, 3,930 respectively. It is worth noting, that this group was the only one that used depth information in their system. The fifth edition (Martínez-Gómez et al, 2014) encouraged participants to use 3D information (point cloud files) with the inclusion of rooms completely imaged in the dark, while also introducing the identification of objects in the scene. The highest result, 6,033.5 points, was obtained by the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China (MIAR ICT) (Xu et al, 2014). They proposed the use of Kernel descriptors for both visual and depth information, while PCA was applied for dimensionality reduction. They used SVM classifiers and managed object recognition and room classification separately. Actually, both problems were expected to be handled together, but none of the participants presented a proposal where the objects appearance (or lack) is used to classify the room.

2.1.5 Lessons Learned

As a general remark, we can point out that, in most editions, the best techniques have been proposed by those participants taking advantage of the introduced novelties. Namely, those proposals that ranked first in the second and third editions were based on the spatial geometry acquired from stereo images. The top performing approach of the fourth edition was the only proposal using range information, and a similar scenario was found in the 2013 edition.

In addition to the generation of solutions to the problem provided by each edition, the Robot Vision task has served for sharing techniques and knowledge between worldwide researchers. This experience has supported several robotic laboratories for generating their own place classifiers, but also for the development of novel approaches that have been successfully deployed in different environments.

Overall, the Robot Vision challenge has provided valuable resources including datasets, benchmarking techniques, and state-of-the-art solutions to the visual place classification problem . Moreover, it has contributed to the generation of a semantic localization researchers community.

2.2 Overview of the Photo Retrieval Task

The Photo Retrieval Task was held over five editions starting in 2007 with the ImageCLEFphoto 2007 Photographic Retrieval Task (Grubinger et al, 2008). The first edition was followed by two very successful years in 2008 (Grubinger et al, 2008) and 2009 (Arni et al, 2009). After this, there was no new edition of the task until 2012 (Zellhöfer, 2012) and 2013 (Zellhöfer, 2013). The main idea of the Photo Retrieval Task was to perform laboratory-style evaluation of visual information retrieval from generic photo collections. In the first three editions of the task, the focus was put on light annotations , text and visual features using generic photo datasets. In the 2012 and 2013 editions, the focus was changed to personalized photo collections following the same principals for annotation and used modalities. The change was made as a consequence of a discussion at ImageCLEF 2011. This seemed not very appealing to participants, and the task did not reach the same number of participants or submissions for these two years, i.e., it was discontinued after 2013. In the following subsections, an overview is provided for the tasks in terms of datasets, metrics used, participants, techniques used and lessons learned.

2.2.1 Datasets

In this subsection, the datasets used are discussed and presented in detail. As mentioned before, the first three editions of the task focused on generic datasets, whereas the two last editions put personalized photo collections into the focus. An overview of all tasks and used datasets can be found in Table 4.

In the 2007 task (Grubinger et al, 2008), 20,000 images were included in the dataset. In addition to the 20,000 color images, the organizers also included several types of metadata . The metadata provided to the participants includes annotation in different languages, title, location, date and additional notes. In addition, participants were given 60 different query topics (structured statements of user needs). The image features provided with the data were rather simple including color histograms, Tamura texture histograms and thumbnails compared with Euclidean distance.

The 2008 task (Arni et al, 2009) changed the focus from just photo retrieval to diversity of the result set, but using the same dataset as in 2007, i.e., the experimental results should show diverse representations of the query. This focus was also continued in the 2009 version of the task, but with a much larger dataset (Lestari Paramita et al, 2010). The dataset for the 2009 task contained almost half a million images (498,920) which was by itself a challenge for the participants.

For both the 2012 (Zellhöfer, 2012) and 2013 (Zellhöfer, 2013) tasks, the dataset changed again. This time the focus was on personal photo collections, and the Pythia dataset containing only 5,555 images was provided. The provided images were uncompressed and taken from 19 different laypersons. The ground truth for the dataset was created using relevance judgments which are highly subjective. From this dataset, the participants could use the following combinations of the provided

Table 4 Details of tasks per year.

Year	Task type	Resource	Images
2007	Photographic ad-hoc retrieval task	IAPR TC-12 Benchmark dataset	20,000
2008	Diversity in Photo Retrieval	IAPR TC-12 Benchmark dataset	20,000
2009	Diversity in Photo Retrieval	Belga dataset	498,920
2012	Personal Photo Retrieval Subtask	Pythia Dataset: uncompressed photographs of 19 laypersons	5,555
2013	Personal Photo Retrieval Subtask	Pythia Dataset: uncompressed photographs of 19 laypersons	5,555

document data and metadata: visual features alone, visual features and metadata, visual features and browsing data, metadata alone, metadata and browsing data, browsing data alone and a combination of all modalities.

2.2.2 Evaluation Measures

For the given tasks, a set of mainly well-known information retrieval metrics were used. This included average precision , precision at rank and F1-measures as harmonic mean of precision and recall . The first two years the tasks mainly focused on precision at a specific rank. To measure diversity, cluster recall was introduced in the 2008 editions of the task which also was used in 2009, but for a different rank and without precision. After a break of two years (2010 and 2011), the new tasks used *normalized Discounted Cumulative Gain (nDCG)* for the evaluation in addition to precision at rank 20 nDCG, which is know to be able to reflect subjectivity and evaluate relevance feedback , was chosen to compensate for the high subjective relevance judgments used to create the dataset ground truth. For the 2013 task, precision at rank 20 was not reported, but instead different cut offs of nDCG (5, 10, 20, 30, 100) and mean average precision with a cut off of 100 were used. Table 5 gives an overview of all metrics used for the different tasks. As one can see, the inconsistency of metrics used for the tasks makes it hard to compare results and measure improvements over the different editions. Nevertheless, since the focus of the task was basically shifted three times (retrieval, diversity, personalized image collections) this does not play an important role.

2.2.3 Participants and Submissions

In general, the Photo Retrieval Task was well received by the community. Especially, the early editions attracted a large number of participants and submissions. In Table 6, the numbers of registered, participated and submitted runs are depicted. Registered indicates the number of people that were interested. Participated shows the number of distinctive teams that submitted a solution whereas submitted runs indicates how many runs the participating teams submitted in total. As mentioned

Table 5 Details of metrics per year.

Year	metrics
2007	mean average precision, precision at rank 20, geometric mean average precision, binary preference
2008	precision at rank 20, cluster recall at rank 20, mean average precision, geometric mean average precision, binary preference, F1-measure
2009	precision at rank 10, cluster recall at rank 10, F1-measure
2012	mean average precision (cut off 100), nDCG‡
2013	nDCG

‡ = discounted cumulative gain of trec_eval v9 with standard discount settings

Table 6 Details of participants and submissions per year.

Year	registered	participated	submitted runs
2007	32	20	616
2008	??	24	1,042
2009	44	19	84
2012 (subtask 1)	64	3†	13†
2012 (subtask 2)	64	2†	8†
2013	10	7	26

†The paper reports that there are excerpts presented in the result tables.

before, the task was very popular in the first three years with a peak of participants and submitted runs of 24 and 1,042 in 2008, respectively. In 2009, only 84 runs were submitted. This was most probably due to the large number of images in the dataset which could be a barrier for teams with not sufficient hardware.

The 2012 and 2013 tasks did not have as many submissions and participants compared to previous years. An explanation for this could not really be found. One reason could be the focus on personalized photo collections and the very subjective ground truth which makes evaluation fuzzy and maybe less interesting to the participants. Furthermore, the number of images in the dataset was also small compared to previous years with 20,000 and 498,920 images.

2.2.4 Techniques used

During different editions of the task, participants used a vast number of different techniques and methods. This is depicted in the number of submitted runs, since for each run, something had to be different compared to a previous one. Overall, 1,789 different runs were submitted in total (most of them for the first three editions). In the following, a summary about the most important aspects throughout all runs is provided.

In 2007 (Grubinger et al, 2008), most participants used the available image annotations for their analysis. The groups submitted 312 bilingual (combination of more than one language) runs and 251 monolingual. 288 runs were concept based (tex-

tual), 276 runs combined text and visual information and only 52 runs used only the visual content features. Most of the runs were based on automatic methods, but a small number also relied on manual approaches (around 3 %). The most used language was English followed by German. The visual content information was the third most used modality.

In 2008 (Arni et al, 2009), the main questions asked for this task were (i) Is it possible to promote diversity withing the top n results?, (ii) Which approaches work best for achieving diversity?, (iii) Does diversity reduce the number of relevant images in the top n results?, (iv) Can text retrieval be used to achieve diverse results?, and (v) How does the performance compare between bilingual and multilingual annotations? The dataset provided was the same as in the 2007 version in terms of images and provided metadata . Overall, 1024 runs were submitted where most of the submissions used the image annotations with 404 runs only using text information. 605 runs used visual information in combination with concept based features. Only 33 runs were purely content based (visual). Comparing the results for mixed, text-only and content-only, the best performance was achieved with mixed (text and visual) followed by text-only and visual-only on the last place. Most of the participants used different re-ranking methods or clustering for the analysis. Apart from that, different ways of merging modalities were applied like combining by scores, etc.

For the 2009 version of the task (Lestari Paramita et al, 2010), the participants were asked to specify the query fields used in their search and the modality of the runs. Query fields were described as T (Title), CT (Cluster Title), CD (Cluster Description) and I (Image). The modality was described as TXT (text-based search only), IMG (content-based image search only) or TXT-IMG (both text and content-based image search). This year, the highest F1 score was different for each modality. A combination of T-CT-I had the highest score in TXT-IMG modality. In the TXT modality, a combination of T-I scored the highest, with T- CT-I following on the second place. However, since only one run used the T-I, it was not enough to provide a conclusion about the best run. Calculating the average F1 score regardless of diversity shows that the best runs are achieved using a combination of Title, Cluster Title and Image. Using all tags in the queries resulted in the worst performance.

None of the participants for the 2012 task (Zellhöfer, 2012) used a combination of all modalities (Zellhöfer, 2012). The participants relied on visual features alone, metadata alone, visual features and metadata, or metadata and browsing data . Interestingly, only one group decided to exploit the browsing data instead of the provided metadata. Surprisingly, they could use this data successfully to solve subtask 1, but reached the last position at subtask 2. This result indicates that there is a particularly strong influence of metadata on the retrieval of events.

Finally, in 2013 (Zellhöfer, 2013), an interesting result of the conducted experiment was that the two leading groups performed almost equally well where one group was relying on sophisticated techniques such as Fisher vectors and local features while the other group used global low-end features embedded in a logical query language. Given the fact, that local features are computationally more inten-

sive than global features, one might further investigate the logical combination of global features in order to achieve comparable results at less computational costs.

2.2.5 Lessons Learned

Based on the information collected from five years of the Photo Retrieval Tasks , several lessons can be learned. The overall conclusions that could be observed in all editions of the task are:

- Multi-modal analysis always improves the performance compared to only text, metadata or visual.
- Diversity is a topic that generates a lot of interest and is seen as important by the community.
- Personalized photo collections are less interesting for the community compared to more generic collections.
- Bilingual retrieval performs nearly as well as monolingual.

A more detailed analysis of the outcome of each task is provided in the following. Comparing the different combinations of provided features showed that using monolingual text achieves the best results followed by bilingual and visual information, respectively. In the monolingual results, Spanish outperforms English and German. In the bilingual results, a combination of English and German achieves the best results. The differences between the different languages are quite small, and the main conclusion was that the query language does only play a small role for the retrieval results. Comparing mixed, text- and visual-only runs, the mixed results (visual + test) outperform the text or visual only results by around 24% on average. Another insight is that manual methods outperform automatic methods, but these are not scalable and therefore unrealistic to use (Grubinger et al, 2008).

The 2008 task holds the record of participants and submitted runs in the series of the photo retrieval task. This is an interesting indicator for the general focus over time on photo retrieval in the community which peaked in 2008 and then decreased. The participants experimented with all different modalities whereas text was still most commonly used. The man insights and lessons learned were that bilingual retrieval performs nearly as well as monolingual (which was also observed in previous years). Combining the concept- and content-based retrieval methods leads to the best results, and the visual retrieval methods got more popular (Arni et al, 2009).

For the 2009 task (Lestari Paramita et al, 2010), the results showed that participants were able to present a diverse result without sacrificing precision. In addition, the results revealed the following insights:

- Information about the cluster title is essential for providing diverse results, as this enables participants to correctly present images based on each cluster. When the cluster information was not being used, the cluster recall score is proven to drop, which showed that participants need better approaches to predict the diversity.
- A combination of title, cluster title and image was proven to maximize the diversity and relevance of the search engine.

- Using mixed modality (text and image) in the runs managed to achieve the highest F1 compared to using only text or image features alone.

Considering the increasing interest of participants in ImageCLEFPhoto, the creation of the new collection was seen as a big achievement in that it provides a more realistic framework for the analysis of diversity and evaluation of retrieval systems aimed at promoting diverse results. The findings from this new collection were found to be promising, and we plan to make use of other diversity algorithms (Dang-Nguyen et al, 2017a) in the future to enable evaluation to be done more thoroughly.

Finally, from the 2012 (Zellhöfer, 2012) and 2013 (Zellhöfer, 2013) tasks, the following insights were gained:

- There was no interest in solving the so-called user-centered initiative of the subtasks. The initiative asked for an alternative representation of the top-k results offering a more diverse view onto the results to the user. This challenge reflects the assumption that a user-centered system should offer users good and varying retrieval results.
- Varying results are likely to compensate for the vagueness inherent in both retrieval and query formulation. Hence, an additional filtering or clustering of the result list could improve the effectiveness and efficiency (in terms of usability) of the retrieval process.
- It remains unclear, if this task was too complex or just out of the area of expertise of the participants that used the dataset for the first time.
- The best performing groups used visual low-level features and metadata to solve the task.
- Again, the utilization of multiple modalities can increase the retrieval effectiveness.

2.3 Overview of the Scalable Image Annotation Tasks

From 2012 to 2016 (Villegas and Paredes, 2012; Villegas et al, 2013; Villegas and Paredes, 2014; Gilbert et al, 2015, 2016), ImageCLEF ran a Scalable Image Annotation task , to promote research into the annotation and classification of images using large-scale and noisy web page data. The primary goal of the challenge was to encourage creative ideas of using web page data to improve image annotation and to develop techniques to allow computers to describe images reliably, localize different concepts depicted and generate descriptions of the scenes. In the 2015 edition (Gilbert et al, 2015), the image annotation task was expanded to concept localization and also natural language sentential description of images. In 2016 edition (Gilbert et al, 2016), the organizers further introduced a text illustration task, to evaluate systems that analyze a text document and select the best illustration for the text from a large collection of images provided .

The challenging issue that was the basis of the image annotation challenges is that every day, users face the ever-increasing quantity of data available to them trying to

find the image on Google of their favorite actress, or the images of the news article someone mentioned at work. Although there are a huge number of images that can be cheaply found from the Internet and a significant amount of information about the image is present on the web pages, the relationship between the surrounding text and images varies greatly, with much of the text being redundant and unrelated. Despite the obvious benefits of using such information in automatic learning, the weak supervision it provides means that it remains a challenging problem.

2.3.1 Datasets

In the first edition (Villegas and Paredes, 2012) of the task, the organizers proposed two subtasks. In the first, the scope was to use both automatically gathered Web data and labeled data to enhance the performance in comparison to using only the labeled data, and in the second, the focus was to use only automatically gathered Web data and language resources to develop a concept scalable annotation system . A training set with 250,000 unlabeled images and textual features and 15,000 images from Flickr, labeled for 94 concepts, was provided to the participants for the first subtask. For subtask 2, the participants were allowed to use only the 250,000 unlabeled images. The test set consisted of 10,000 labeled images for the same 94 concepts of the training set and 2,000 labeled images for 105 concepts respectively. In 2013 and 2014 (Villegas et al, 2013; Villegas and Paredes, 2014), only one task was proposed whose purpose was developing concept scalable image annotation systems using only automatically gathered Web data. In these editions participants were provided with 250,000 Web images and respective Web-pages in 2013 and 500,000 images and respective Web-pages in 2014. In the second edition, the development set was composed by 1,000 labeled images for 95 concepts, but in the test set, there were 2,000 images, and the participants had to label them for 116 concepts. In 2014, the participants had to label 7,291 samples for 207 concepts, 100 unseen in development (see Table 7). In 2015 and 2016 (Gilbert et al, 2015, 2016), the participants were provided with unlabeled Web images, their respective Web-pages, and textual features. In the fourth edition, two subtasks were proposed, the first was related to image annotation as usual adding also a localization requirement. In the second, a completely new task, the participants were requested to develop a system that could describe an image with a textual description of the visual content depicted in the image. The development set contained 1,979 and the test set 3,070 labeled images. In a second track (“clean track”) of the second subtask participants were provided with a test set of 450 images with bounding boxes labeled with concepts. Both development set and test sets were subsets of the 500,000 training images. In 2016, the three subtasks remained the same (the “clean track” became a subtask) but the organizers increased the number of images in the training set (510,000) and in the development (2,000). In addition, a ”teaser” task was proposed where participants were asked to analyze a given text document and find the best illustration for it from a set of all available images. The training set consisted of approximately 300,000 documents from the entire corpus. The remaining 200,000 have been used for testing. A sepa-

Table 7 Number of concepts and training, development and test set.

Task edition	Number of concepts	Training images	Development images	Test images
2012 (subtask 1)	94	250,000 + 15,000	-	10,000
2012 (subtask 2)	105	250,000	1,000	2,000
2013	116	250,000	1,000	2,000
2014	207	500,000	1,940	7,291
2015	251	500,000	1,979	3,070
2016	251	510,000	2,000	3,070
2016 ('teaser' task)	251	310,000	3,000	200,000

rate development set of about 3,000 image-webpage pairs were also provided as a validation set for parameter tuning and optimization purposes.

2.3.2 Evaluation Measures

The performance measures have been computed for each of the test set images, and as a global measure, the mean of these measures has been obtained. The measures used from 2012 to 2014 for comparing the submitted systems were *Average Precision (AP)* and *F-measure*. In addition in 2012, the *Interpolated Average Precision (IAP)* was also used.

$$AP = \frac{1}{C}\sum_{c=1}^{C}\frac{c}{rank(c)} \tag{1}$$

$$IAP = \frac{1}{C}\sum_{c=1}^{C}\max_{c'\geq c}\frac{c'}{rank(c')} \tag{2}$$

$$F = 2\cdot\frac{precision\cdot recall}{precision + recall} \tag{3}$$

where 'C' is the number of ground truth concepts for the image, 'rank(c)' is the rank position of the c-th ranked ground truth concept, and 'precision' and 'recall' are respectively the precision and recall for the annotation decisions. AP and IAP depended only on the confidence scores, while the F-measure only depended on the annotation decisions given by participants to the images.

Since the number of concepts per image was small and variable, the AP and IAP have been computed using the rank positions (precision = c/rank(c)), i.e., using every possible value of recall, instead of using some fixed values of recall. After obtaining the means, these measures can be referred to as: *Mean Average Precision (MAP)*, *Mean Interpolated Average Precision (MIAP)* and *Mean F-measure (MF)*, respectively.

In 2015 and 2016, the localization of subtasks 1 have been 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-

Table 8 Participation in the Scalable Image Annotation Tasks over the years.

Participation	2012	2013	2014	2015	2016
Registered groups	55	104	43	154	82
Participant groups	3	13	11	14	7
Working notes submitted	2	9	9	11	7

truth segmentation, divided by the area of their union. The final results have been presented both in terms of average performance over all images of all concepts, and also per concept performance over all images. Subtask 2 has been evaluated using the Meteor[1] evaluation metric against a minimum of five human-authored textual descriptions. Systems participating in the "clean track" (called subtask 3 in 2016) have additionally had the option of being evaluated with a fine-grained metric, which was the average F-measure across all test images on how well the text generation system selected the correct concepts to be described (against the ground truth). In 2016, in addition, for the 'teaser' task, the test images have been ranked according to their distance to the query article. Recall at the k-th rank position (R@K) of the ground truth image have been used as performance metrics. Several values of k have been used, and participants were asked to submit the top 100 ranked images.

2.3.3 Participants and Submissions

The Scalable Image Annotation tasks , over the years, have had changing fortunes. After a somewhat weak start in 2012, where compared to 55 registered groups that signed the license agreement and therefore had access for downloading the datasets, only 26 runs were submitted for three groups on the two subtasks, the number of participants increased up to 2015. In 2013 and 2014, 58 runs were submitted for 13 and 11 groups, respectively, increasing to 122 runs submitted in 2015 for 14 groups from all parts of the world including China, France, Tunisia, Colombia, Japan, Romania. In 2016, unfortunately, the participation was not as good as in previous years. In total, 82 groups signed the license agreement, but only seven groups took part in the task and submitted 50 system runs overall (see Table 8).

2.3.4 Techniques used

The first attempts to participate in the Scalable Image Annotation challenge were based mainly on scalability (Ushiku et al, 2012b) using a combination of several SIFT features. For annotation, they used an online learning method *Passive-Aggressive with Averaged Pairwise Loss (PAAPL)* and labeled the Web-data using the appearance of concept words in the textual features.

[1] http://www.cs.cmu.edu/ alavie/METEOR/

The following years, the participation was much higher. Most of the submitted runs significantly outperformed the baseline system, but very large differences can be observed amongst the systems. In 2013, for both MAP and MF, the improvement was from below 10% to over 40%. An interesting detail to note is that for MAP there were several top performing systems. However, when comparing to the respective MF measures, the *CNRS TELECOM ParisTech (TPT)* submissions (Sahbi, 2013) clearly outperform the rest. The key difference between these was the method for deciding which concepts were selected for a given image. This leads us to believe that many of the approaches could be improved greatly by changing that last step of their systems. Many of the participants have chosen to use the same scheme as the baseline system proposed by organizers for selecting the concepts, i.e., the top N and fixed for all images. The number of concepts per image was expected to be variable, thus making this strategy less than optimal. In contrast to usual image annotation evaluations with labeled training data, this challenge required facing different problems, such as handling the noisy data, textual processing and multi-label annotations. This permitted the participants to concentrate their efforts in different aspects. Several teams extracted their own visual features , for which they observed improvements with respect to the features provided by the organizers. For the textual processing, several different approaches were tried by the participants and some of these teams such as MIL (Hidaka et al, 2013), UNIMORE (Grana et al, 2013), CEA LIST (Borgne et al, 2013), and URJC&UNED (Sánchez-Oro et al, 2013) reported that as more information and additional resources are used the performance of the systems improved.

After the first appearance in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012 (Krizhevsky et al, 2012), *Convolutional Neural Network (CNN)* gained great popularity for its good performances on many classification tasks. Thus, it is no wonder that in 2014, three groups based their system on CNN pre-trained using ImageNet. Two of the teams, MIL (Kanehira et al, 2014) and MindLab (Vanegas et al, 2014), used the CNN output of an intermediate layer as a visual feature. In (Krizhevsky et al, 2012), it has been shown also that features extracted from the upper layers of the CNN can also serve as good descriptors for image retrieval . It implies that a CNN trained for a specific task has acquired generic representation of objects that will be useful for all sorts of visual recognition tasks (Babenko et al, 2014; Piras and Giacinto, 2017). The third team that used CNN was MLIA (Xu et al, 2014), which employed the synsets predicted by the CNN to clean the concepts automatically assigned using the Web page data. In this case, the performance of the system could be greatly affected if the concepts for annotation differ significantly from the ones of ImageNet. As in previous years, most of the teams proposed approaches based on trained classifiers. In the case of the MIL team, the classifier is multi-label so each time the list of concepts to detect changed, the classifier had to be retrained. However, the PAAPL (Ushiku et al, 2012a) algorithm of MIL is designed with special consideration of scalability, so in their case, it does not seem an issue. Another approach that has been adopted was the use of one classifier per concept, trained one concept at a time using positive and negative samples. For scalability, the learning should be based on a selection of negative im-

ages so that this process is independent of how many concepts there are. Although many groups found it adequate, with respect to a multi-label classifier this might not be the optimal approach.

In the fourth edition of the challenge, the requirement of labeling and localizing all 500,000 images was added. Three groups achieved over 0.50 MAP across the evaluation set with 50% overlap with the ground-truth. This is an excellent result given the challenging nature of the images used and the wide range of concepts provided. Also in this edition, CNNs have been the masters and allowed for large improvements in performance. All of the top 4 groups used CNNs in their pipeline for the feature description. *Social Media and Internet Vision Analytics Lab (SMIVA)* (Kakar et al, 2015) used a deep learning framework with additional annotated data, while IVANLPR (Li et al, 2015b) 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. RUC (Li et al, 2015a) trained per concept, an ensemble of linear SVMs trained by Negative Bootstrap using CNN features as image representation. Concept localization was achieved by classifying object proposals generated by Selective Search. The approach by CEA LIST (Gadeski et al, 2015) was very simple, they just use the CNN features in a small grid based approach for localization. In this edition, two other subtasks were proposed. For subtask 2, participants were asked to generate sentence-level textual descriptions for all 500,000 training images and the systems were evaluated on a subset of 3,070 instances. RUC (Li et al, 2015a) used the state-of-the-art deep learning based *CNN Long Short-Term Memory Network (CNN-LSTM)* caption generation system, MindLab (Pellegrin et al, 2015) employed a joint image-text retrieval approach, and UAIC (Calfa et al, 2015) a template-based approach. Two of the teams that participated in this subtask participated also in the subtask 3 where participants were asked to generate textual descriptions for 450 test images based on labeled bounding box input. RUC (Li et al, 2015a) used a deep learning based sentence generator coupled with re-ranking based on the bounding box input, while UAIC (Calfa et al, 2015) used a template-based generator.

In the 2016 edition, subtask 1 had a lower participation than in the previous year. However, there were some excellent results showing improvements over previous editions. CEA LIST (Borgne et al, 2016) used a deep learning framework (Simonyan and Zisserman, 2014), but focused on improving the localization of the concepts. They attempted to use a face body part detector, boosted by previous years results. MRIM-LIG (Portaz et al, 2016) also used a classical deep learning framework and the object localization (Uijlings et al, 2013), where an *a priori* set of bounding boxes were defined which were expected to contain a single concept each. CNRS (Sahbi, 2016) focused on concept detection and used label enrichment to increase the training data quantity in conjunction with an SVM and VGG (Simonyan and Zisserman, 2014) deep network.

2.3.5 Lessons Learned

The Scalable Image Annotation challenge had the objective of taking advantage of automatically gathered image and textual Web data for training, in order to develop more scalable image annotation systems. Even if in the subtask 1 of the first edition none of the participants were able to use the web-data to obtain a better performance than when using only manually labeled data, in the subtask 2, the results were somewhat positive. In some cases, the performance was even comparable to good annotation systems learned using manually labeled data. Over the years, some groups participated several times (e.g., MIL, KDEVIR, CEA LIST, TPT, INAOE), and in every edition, they were able to improve the results obtained in the previous one in particular improving more for the MF measures. This indicates a greater success in the developed techniques for choosing the final annotated concepts. In 2015, the requirement of labeling and localizing all 500,000 images was introduced. However, a limitation in the dataset has arisen: the difficulty of ensuring the ground truth has 100% of concepts labeled. This was especially problematic as the concepts selected include fine-grained categories such as eyes and hands that are generally small but occur frequently in the dataset. In addition, it was difficult for annotators to reach a consensus in annotating bounding boxes for less well-defined categories such as trees and field. Another interesting aspect of this challenge that has been going on for a long time is that the increased CNN usage as the feature representation had improved localization techniques and the performances have been progressively improved even under this point of view.

2.4 Overview of the Lifelog Tasks

The main goal of the Lifelog task is to advance the state-of-the-art research in lifelogging as an application of information retrieval. To do this, for each edition, a standard dataset was provided and together with the dataset, and tasks were introduced. In the first edition, ImageCLEFlifelog2017 (Dang-Nguyen et al, 2017b), *Lifelog Retrieval Task (LRT)* and *Lifelog Summarization Task (LST)* were introduced, while in the second edition, ImageCLEFlifelog2018 (Dang-Nguyen et al, 2018), the LRT task was improved and renamed as *Activities of Daily Living (ADL)* understanding task. The details of these tasks are:

- *Lifelog Retrieval Task (LRT)*. In this task, the participants had to analyse the lifelog data and for several specific queries, return the correct answers. For example: *"In a Meeting: Find the moment(s) in which the user was in a meeting at work with 2 or more people. To be considered relevant, the moment must occur at meeting room and must contain at least two colleagues sitting around a table at the meeting. Meetings that occur outside of my place of work are not relevant."*
- *Lifelog Summarization Task (LST)*. In this task, the participants had to analyse all the images and summarize them according to specific requirements. For instance: *"Shopping: Summarize the moment(s) in which user doing shopping. To*

be relevant, the user must clearly be inside a supermarket or shopping stores (includes book store, convenient store, pharmacy, etc). Passing by or otherwise seeing a supermarket are not considered relevant if the user does not enter the shop to go shopping. Blurred or out of focus images are not relevant. Images that are covered (mostly by the lifelogger's arm) are not relevant." In this task, not only the relevance is considered, but participants are also asked to provide the diversification of the selected images with respect to the target scenario.

- *Activities of Daily Living (ADL)* understanding task: For this task, given a period of time, e.g., *"From 13 August to 16 August"* or *"Every Saturday"*, the participants should analyse the lifelog data and provide a summarisation based on the selected concepts (provided by the task organizers) of ADL and the environmental settings/ contexts in which these activities take place. Some examples of ADL concepts: *"Commuting (to work or other common venue)"*, *"Travelling (to a destination other than work, home or some other common social event)"*, and contexts: *"In an office environment"*, *"In a home"*, *"In an open space"*. The summarisation should be described as the frequency and time spent for ADL concepts and total time for contexts concepts. For example: ADL: *"Eating/drinking: 6 times, 90 minutes"*, *"Travelling: 1 time, 60 minutes"*; context: *"In an office environment: 500 minutes"*, *"In a church: 30 minutes"*.

2.4.1 Datasets

In the first edition, ImageCLEFlifelog2017, the dataset was developed based on the dataset in *National Institute of Informatics (NII) Testbeds and Community for Information access Research (NTCIR)*-12 (Gurrin et al, 2016) . This dataset consists of data from three lifeloggers for a period of about one month each. The data contains 88,124 wearable camera images (about 1,500 - 2,500 images per day), an *eXtensible Markup Language (XML)* description of 130 associated semantic locations (e.g., Starbucks cafe, McDonalds restaurant, home, work) and the four physical activities, detected by the Moves app[2] installed in the lifeloggers's phone: walking, cycling, running and transport of the lifeloggers at a granularity of one minute. Together with the locations, activities and visual concepts are provided as the output of the Caffe CNN-based visual concept detector (Jia et al, 2014). This classifier provided labels and probabilities for 1,000 objects in every image. The accuracy of the Caffe visual concept detector is variable and is representative of the current generation of off-the-shelf visual analytics tools.

In the second edition, ImageCLEFlifelog2018, the dataset was developed based on the dataset in NTCIR-13 (Gurrin et al, 2017) . This dataset contains richer data with respect to the previous edition, where more biometrics information was added. The visual concept was also improved by using Microsoft Computer Vision API[3].

[2] http://moves-app.com/

[3] https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/

Participants were provided with two different sets of topics: the development set (devset) for developing and training their methods and the test set (testset) for the final evaluation. A summary of the data collection is shown in Table 9.

Table 9 Statistics of the Lifelog dataset.

Year	2017	2018
Number of days	90	50
Size of the dataset (GB)	18.18	18.85
Number of images	88,124	88,440
Number of locations	130	135
Biometrics information	No	Yes
Visual concepts	Caffe concepts	Microsoft CV API
Personal annotation	No	Yes
Music information	No	Yes
Number of LRT topics	36	-
Number of LST topics	15	20
Number of ADL topics	-	20

2.4.2 Evaluation Measures

For the LRT, evaluation metrics based on nDCG at different depths, i.e., nDCG@N, were used. In this task, N was chosen from $\{5,10\}$, depending on the topics. In the LST, 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., $X = 5,10,20,30,40,50$. The official ranking metrics this year was the **F1-measure@10** or images, which gives equal importance to diversity (via $CR@10$) and relevance (via $P@10$).

In the ADL understanding task, the score is computed as as follows:

$$ADL_{score} = \frac{1}{2}\left(max\{0,1-\frac{|n-n_{gt}|}{n_{gt}}\}+max\{0,1-\frac{|m-m_{gt}|}{m_{gt}}\}\right)$$

where n, n_{gt} are the submitted and ground-truth values for how many times the events occurred, respectively, and m, m_{gt} are the submitted and ground-truth values for how long the events happened, respectively.

2.4.3 Participants and Submissions

In the first edition challenging task, eleven teams were registered, of which three teams took part in the task and submitted overall 19 runs. All three participating teams submitted a working paper describing their system.

In the second run, the number of participants was considerably higher compared to 2017 with 25 registered teams which submitted in total 41 runs: 29 (21 official, 8 additional) for LMRT and 12 (8 official, 4 additional) for ADLT, from 7 teams from Brunei, Taiwan, Vietnam, Greece-Spain, Tunisia, Romania, and a multi-nation team from Ireland, Italy, Austria, and Norway. The approaches employed ranged from fully automatic to fully manual, from using a single information source provided by the task to using all information as well as integrating additional resources, from traditional learning methods (e.g., SVMs) to deep learning and ad-hoc rules.

Table 10 Details of participants and submissions per year of the Lifelog task.

Year	Registered	Participated	Submitted runs
2017	11	3	19
2018	25	7	41

2.4.4 Techniques used

In the first edition, most of the teams proposed to only explore the visual or combine visual and textual information. In (Molino et al, 2017), the authors proposed a three-step method as follow: As a first step, they filtered out images with very homogeneous colors and with a high blurriness. Then, the system ranked the remaining images and clustered the top ranked images into a series of events using either k-means or a hierarchical tree. Finally, they selected, in an iterative manner, as many images per cluster as to fill a fixed size bucket (50 as required by the tasks). The study in (Dogariu and Ionescu, 2017) proposed an approach that combines textual and visual information in the process of selecting the best candidates for the tasks requirements. The run that they submitted relied solely on the information provided by the organizers and no additional annotations or external data, nor feedback from the users had been employed. Additionally, a multi-stage approach has been used. The algorithm starts by analyzing the concept detectors output provided by the organizers and selecting for each image only the most probable concepts. From the list of the topics, each of them has then been parsed such that only relevant words have been kept and information regarding location, activity and the targeted user are extracted as well. The images that did not fit the topic requirements have been removed and this shortlist of images is then subject to a clustering step . Finally, the results are pruned with the help of similarity scores computed using WordNets built-in similarity distance functions.

Learning from the drawbacks of the first edition, most of the teams participating in the second edition exploited multi-modal data by combining visual, text, location and other information to solve the tasks, which is different from the previous year when often only one type of data was analyzed. Furthermore, deep learning was exploited by many teams (Tran et al, 2018; Dogariu and Ionescu, 2018; Abdallah et al, 2018). For example, in (Dogariu and Ionescu, 2018), CAMPUS-UPB team extracted the visual concepts using a CNN approach and then combined the extracted features with other information and clustered them using K-means and reranked using the concepts and queried topics. In the method proposed by the Regim Lab team (Abdallah et al, 2018), combinations of visual features , textual features and a combination of both by XQuery FLOWR, then fine tuned by CNN architectures were used. For the visual features fine tuned CNN architectures were utilized. Beside exploiting multi-modal data and deep learning , natural language processing was also considered . NLP-Lab (Tang et al, 2018) team reduced user involvement during the retrieval by using natural language processing. In this method, visual concepts were extracted from the images and combined with textual knowledge to get rid of the noise. For ADL, the images are ranked by time and frequency, whereas for LRT ranking is performed exploiting similarity between image concepts and user queries.

Different from the competitive teams, the task organizer team proposed only baseline approaches, with the purpose to serve as referent results for the participants. In the first edition (Zhou et al, 2017), they proposed multiple approaches, from fully automatic to fully manual paradigm. These approaches started by grouping similar moments together based on time and concepts. By applying this chronological-based segmentation, the problem of image retrieval turned into image segments retrieval. Starting from a topic query, it is transformed into small queries where each is asking for a single piece of information of concepts, location, activity, and time. The moments that matched all of those requirements are returned as the retrieval results. In order to remove the non-relevant images, a filtering step is applied on the retrieved images, by removing blurred images and images that are mainly covered by a huge object or by the arms of the user. Finally, the images are diversified into clusters and the top images that close to center are selected for the summarization, which can be done automatically or using *Relevance Feedback (RF)*. In the second edition, the organizers proposed an improved version of the baseline search engine (Zhou et al, 2018), named LIFER, and based on that, an interactive lifelog search interface was built allowing users to solve both tasks in the competition.

2.4.5 Lessons Learned

We learned that in order to retrieve moments from lifelog data efficiently, we should exploit and combine multi-modal information, from visual, textual, location information to biometrics and the usage data from the lifeloggers devices. Furthermore, we learned that lifelogging is following the trend in data analytics, meaning that deep learning is being exploited in most of the methods. We also learned that there

is still room for improvement, since the best results are coming from the fine-tuned queries, which means we need more advanced techniques for bridging the gap between an abstraction of human needs and the multi-modal data.

Regarding the participants, the significant improvement of the second edition compared to the first one shows how interesting and challenging lifelog data is and that it holds much research potential.

All in all, the task was quite successful for the first two runs, tacking into account that lifelogging is a rather new and not common field. The tasks helped to raise more awareness for lifelogging in general, but also to point at the potential research questions such as the previous mentioned multi-modal analysis, system aspects for efficiency, etc.

As next steps, the organizers do not plan to enrich the dataset but rather provide richer data and narrow down the applications of the challenges (e.g., extend to health-care application).

3 Discussion and Conclusions

From the descriptions of the different image retrieval tasks , some overall lessons and insights can be gained. Specifically the following insights are the most important:

- The consistent discrepancy between the registered groups and those eventually participating in the various challenges is a clear sign of interest in the data, and perhaps even more into the evaluation measures and experimental protocols developed over the years by all organizers. This is a testament to the ability of ImageCLEF of influencing and steering research in the community towards challenging goals.
- All tasks have significantly evolved over their lifetime, managing a fine balance between building a core community of participants that could leverage over prior experience in participating in the tasks, and continuously pushing the envelope in proposing new, cutting edge challenges supporting timely research in the respective field. This is witnessed by the popularity of the data and setup developed over the years.
- Lastly, the fundamental vision behind ImageCLEF has not been particularly affected by the deep learning tidal wave that did hit the community in the last years. On the contrary, ImageCLEF has continued thriving during this paradigm shift, and in several circumstances, it has been able to take advantage of it.

In conclusion, over its very long lifetime, ImageCLEF has been consistently a firm reference point in visual benchmarking and reproducibility, providing resources, promoting fundamental research questions and overall contributing strongly to the quest for intelligent seeing machines.

References

Abdallah FB, Feki G, Ezzarka M, Ammar AB, Amar CB (2018) Regim Lab Team at ImageCLEF-lifelog LMRT Task 2018. In: (Cappellato et al, 2018)

Agosti M, Ferro N, Peters C, de Rijke M, Smeaton A (eds) (2010) Multilingual and Multimodal Information Access Evaluation. Proceedings of the International Conference of the Cross-Language Evaluation Forum (CLEF 2010), Lecture Notes in Computer Science (LNCS) 6360, Springer, Heidelberg, Germany

Arni T, Clough P, Sanderson M, Grubinger M (2009) Overview of the ImageCLEFphoto 2008 Photographic Retrieval Task. In: (Peters et al, 2009), pp 500–511

Babenko A, Slesarev A, Chigorin A, Lempitsky VS (2014) Neural codes for image retrieval. In: Fleet DJ, Pajdla T, Schiele B, Tuytelaars T (eds) Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I, Springer, Lecture Notes in Computer Science, vol 8689, pp 584–599, DOI 10.1007/978-3-319-10590-1, URL https://doi.org/10.1007/978-3-319-10590-1

Balog K, Cappellato L, Ferro N, Macdonald C (eds) (2016) CLEF 2016 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1609/

Borgne HL, Popescu A, Znaidia A (2013) CEA list@imageclef 2013: Scalable concept image annotation. In: (Forner et al, 2013)

Borgne HL, Gadeski E, Chami I, Tran TQN, Tamaazousti Y, Gînsca A, Popescu A (2016) Image annotation and two paths to text illustration. In: (Balog et al, 2016)

Braschler M, Harman DK, Pianta E, Ferro N (eds) (2010) CLEF 2010 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1176/

Calfa A, Silion D, Bursuc AC, Acatrinei CP, Lupu RI, Cozma AE, Padurariu C, Iftene A (2015) Using textual and visual processing in scalable concept image annotation challenge. In: (Cappellato et al, 2015)

Cappellato L, Ferro N, Halvey M, Kraaij W (eds) (2014) CLEF 2014 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1180/

Cappellato L, Ferro N, Jones GJF, SanJuan E (eds) (2015) CLEF 2015 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1391/

Cappellato L, Ferro N, Goeuriot L, Mandl T (eds) (2017) CLEF 2017 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1866/

Cappellato L, Ferro N, Nie JY, Soulier L (eds) (2018) CLEF 2018 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-2125/

Clough P, Sanderson M (2004) The CLEF 2003 Cross Language Image Retrieval Track. In: Peters C, Braschler M, Gonzalo J, Kluck M (eds) Comparative Evaluation of Multilingual Information Access Systems: Fourth Workshop of the Cross–Language Evaluation Forum (CLEF 2003) Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 3237, Springer, Heidelberg, Germany, pp 581–593

Clough P, Müller H, Sanderson M (2005) The CLEF 2004 Cross-Language Image Retrieval Track. In: Peters C, Clough P, Gonzalo J, Jones GJF, Kluck M, Magnini B (eds) Multilingual Information Access for Text, Speech and Images: Fifth Workshop of the Cross–Language Evaluation Forum (CLEF 2004) Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 3491, Springer, Heidelberg, Germany, pp 597–613

Clough P, Müller H, Deselaers T, Grubinger M, Lehmann TM, Jensen J, Hersh W (2006) The CLEF 2005 Cross–Language Image Retrieval Track. In: Peters C, Gey FC, Gonzalo J, Jones GJF, Kluck M, Magnini B, Müller H, de Rijke M (eds) Accessing Multilingual Information

Repositories: Sixth Workshop of the Cross–Language Evaluation Forum (CLEF 2005). Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 4022, Springer, Heidelberg, Germany, pp 535–557

Clough P, Grubinger M, Deselaers T, Hanbury A, Müller H (2007) Overview of the ImageCLEF 2006 Photographic Retrieval and Object Annotation Tasks. In: Peters C, Clough P, Gey FC, Karlgren J, Magnini B, Oard DW, de Rijke M, Stempfhuber M (eds) Evaluation of Multilingual and Multi-modal Information Retrieval : Seventh Workshop of the Cross–Language Evaluation Forum (CLEF 2006). Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 4730, Springer, Heidelberg, Germany, pp 223–256

Dang-Nguyen DT, Piras L, Giacinto G, Boato G, Natale FGBD (2017a) Multimodal retrieval with diversification and relevance feedback for tourist attraction images. ACM Trans Multimedia Comput Commun Appl 13(4):49:1–49:24, DOI 10.1145/3103613, URL http://doi.acm.org/10.1145/3103613

Dang-Nguyen DT, Piras L, Riegler M, Boato G, Zhou L, Gurrin C (2017b) Overview of ImageCLEFlifelog 2017: Lifelog Retrieval and Summarization. In: (Cappellato et al, 2017)

Dang-Nguyen DT, Piras L, Riegler M, Zhou L, Lux M, Gurrin C (2018) Overview of ImageCLEFlifelog 2018: Daily Living Understanding and Lifelog Moment Retrieval. In: (Cappellato et al, 2018)

Deselaers T, Hanbury A (2009) The Visual Concept Detection Task in ImageCLEF 2008. In: (Peters et al, 2009), pp 531–538

Deselaers T, Hanbury A, Viitaniemi V, Benczúr AA, Brendel M, Daróczy B, Escalante Balderas HJ, Gevers T, Hernández-Gracidas CA, Hoi SCH, Laaksonen J, Li M, Marín Castro HM, Ney H, Rui X, Sebe N, Stöttinger J, Wu L (2008) Overview of the ImageCLEF 2007 Object Retrieval Task. In: (Peters et al, 2008), pp 445–471

Dogariu M, Ionescu B (2017) A Textual Filtering of HOG-based Hierarchical Clustering of Lifelog Data. In: (Cappellato et al, 2017)

Dogariu M, Ionescu B (2018) Multimedia Lab @ CAMPUS at ImageCLEFlifelog 2018 Lifelog Moment Retrieval. In: (Cappellato et al, 2018)

Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) (2012) CLEF 2012 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1178/

Forner P, Navigli R, Tufis D, Ferro N (eds) (2013) CLEF 2013 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1179/

Fornoni M, Martínez-Gómez J, Caputo B (2010) A multi cue discriminative approach to semantic place classification. In: (Agosti et al, 2010)

Fraundorfer F, Wu C, Pollefeys M (2010) Methods for combined monocular and stereo mobile robot localization. In: (Ünay et al, 2010), pp 180–189, DOI 10.1007/978-3-642-17711-8_19, URL https://doi.org/10.1007/978-3-642-17711-8_19

Gadeski E, Borgne HL, Popescu A (2015) CEA list's participation to the scalable concept image annotation task of imageclef 2015. In: (Cappellato et al, 2015)

Gilbert A, Piras L, Wang J, Yan F, Dellandréa E, Gaizauskas RJ, Villegas M, Mikolajczyk K (2015) Overview of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task. In: (Cappellato et al, 2015)

Gilbert A, Piras L, Wang J, Yan F, Ramisa A, Dellandréa E, Gaizauskas RJ, Villegas M, Mikolajczyk K (2016) Overview of the ImageCLEF 2016 Scalable Concept Image Annotation Task. In: (Balog et al, 2016)

Grana C, Serra G, Manfredi M, Cucchiara R, Martoglia R, Mandreoli F (2013) UNIMORE at imageclef 2013: Scalable concept image annotation. In: (Forner et al, 2013)

Grubinger M, Clough P, Hanbury A, Müller H (2008) Overview of the ImageCLEFphoto 2007 Photographic Retrieval Task. In: (Peters et al, 2008), pp 433–444

Gurrin C, Joho H, Hopfgartner F, Zhou L, Albatal R (2016) NTCIR Lifelog: The First Test Collection for Lifelog Research. In: Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, pp 705–708

Gurrin C, Joho H, Hopfgartner F, Zhou L, Gupta R, Albatal R, Dang-Nguyen DT (2017) Overview of NTCIR-13 Lifelog-2 Task. In: Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies

Hidaka M, Gunji N, Harada T (2013) MIL at imageclef 2013: Scalable system for image annotation. In: (Forner et al, 2013)

Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22Nd ACM International Conference on Multimedia, ACM, New York, NY, USA, MM '14, pp 675–678, DOI 10.1145/2647868.2654889, URL http://doi.acm.org/10.1145/2647868.2654889

Kakar P, Wang X, Chia AY (2015) Automatic image annotation using weakly labelled web data. In: (Cappellato et al, 2015)

Kanehira A, Hidaka M, Mukuta Y, Tsuchiya Y, Mano T, Harada T (2014) MIL at imageclef 2014: Scalable system for image annotation. In: (Cappellato et al, 2014)

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Bartlett PL, Pereira FCN, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States., pp 1106–1114, URL http://papers.nips.cc/book/advances-in-neural-information-processing-systems-25-2012

Lestari Paramita M, Sanderson M, Clough P (2010) Diversity in Photo Retrieval: Overview of the ImageCLEFPhoto Task 2009. In: (Peters et al, 2010), pp 45–59

Li X, Jin Q, Liao S, Liang J, He X, Huo Y, Lan W, Xiao B, Lu Y, Xu J (2015a) Ruc-tencent at imageclef 2015: Concept detection, localization and sentence generation. In: (Cappellato et al, 2015)

Li Y, Liu J, Wang Y, Liu B, Fu J, Gao Y, Wu H, Song H, Ying P, Lu H (2015b) Hybrid learning framework for large-scale web image annotation and localization. In: (Cappellato et al, 2015)

Luo J, Pronobis A, Caputo B, Jensfelt P (2007) Incremental learning for place recognition in dynamic environments. In: 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 29 - November 2, 2007, Sheraton Hotel and Marina, San Diego, California, USA, pp 721–728

Martínez-Gómez J, Jiménez-Picazo A, García-Varea I (2009) A particle-filter based self-localization method using invariant features as visual information. In: Peters C, Ferro N (eds) Working Notes for CLEF 2009 Workshop co-located with the 13th European Conference on Digital Libraries (ECDL 2009) , Corfù, Greece, September 30 - October 2, 2009., CEUR-WS.org, CEUR Workshop Proceedings, vol 1175, URL http://ceur-ws.org/Vol-1175/CLEF2009wn-ImageCLEF-MartinezGomezEt2009.pdf

Martínez-Gómez J, Jiménez-Picazo A, Gámez JA, García-Varea I (2010) Combining image invariant features and clustering techniques for visual place classification. In: (Ünay et al, 2010), pp 200–209, DOI 10.1007/978-3-642-17711-8_21, URL https://doi.org/10.1007/978-3-642-17711-8_21

Martínez-Gómez J, García-Varea I, Caputo B (2012) Overview of the ImageCLEF 2012 Robot Vision Task. In: (Forner et al, 2012)

Martínez-Gómez J, García-Varea I, Cazorla M, Caputo B (2013) Overview of the ImageCLEF 2013 Robot Vision Task. In: (Forner et al, 2013)

Martínez-Gómez J, García-Varea I, Cazorla M, Morell V (2014) Overview of the ImageCLEF 2014 Robot Vision Task. In: (Cappellato et al, 2014)

Molino AGD, Mandal B, Lin J, Lim JH, Subbaraju V, Chandrasekhar V (2017) VC-I2R@ImageCLEF2017: Ensemble of Deep Learned Features for Lifelog Video Summarization. In: (Cappellato et al, 2017)

Nilsback M, Caputo B (2004) Cue integration through discriminative accumulation. In: 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), with CD-ROM, 27 June - 2 July 2004, Washington, DC, USA, IEEE Computer Society, pp 578–

585, DOI 10.1109/CVPR.2004.67, URL http://doi.ieeecomputersociety.org/10.1109/CVPR.2004.67

Nowak S, Dunker P (2010) Overview of the CLEF 2009 Large-Scale Visual Concept Detection and Annotation Task. In: (Peters et al, 2010), pp 94–109

Nowak S, Huiskes MJ (2010) New Strategies for Image Annotation: Overview of the Photo Annotation Task at ImageCLEF 2010. In: (Braschler et al, 2010)

Nowak S, Nagel K, Liebetrau J (2011) The CLEF 2011 Photo Annotation and Concept-based Retrieval Tasks. In: Petras V, Forner P, Clough P, Ferro N (eds) CLEF 2011 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1177/

Pellegrin L, Vanegas JA, Ovalle JEA, Beltrán V, Escalante HJ, Montes-y-Gómez M, González FA (2015) INAOE-UNAL at imageclef 2015: Scalable concept image annotation. In: (Cappellato et al, 2015)

Peters C, Jijkoun V, Mandl T, Müller H, Oard DW, Peñas A, Petras V, Santos D (eds) (2008) Advances in Multilingual and Multimodal Information Retrieval: Eighth Workshop of the Cross–Language Evaluation Forum (CLEF 2007). Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 5152, Springer, Heidelberg, Germany

Peters C, Deselaers T, Ferro N, Gonzalo J, Jones GJF, Kurimo M, Mandl T, Peñas A (eds) (2009) Evaluating Systems for Multilingual and Multimodal Information Access: Ninth Workshop of the Cross–Language Evaluation Forum (CLEF 2008). Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 5706, Springer, Heidelberg, Germany

Peters C, Tsikrika T, Müller H, Kalpathy-Cramer J, Jones GJF, Gonzalo J, Caputo B (eds) (2010) Multilingual Information Access Evaluation Vol. II Multimedia Experiments – Tenth Workshop of the Cross–Language Evaluation Forum (CLEF 2009). Revised Selected Papers, Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany

Piras L, Giacinto G (2017) Information fusion in content based image retrieval: A comprehensive overview. Information Fusion 37:50 – 60

Portaz M, Budnik M, Mulhem P, Poignant J (2016) MRIM-LIG at imageclef 2016 scalable concept image annotation task. In: (Balog et al, 2016)

Pronobis A, Caputo B (2009) COLD: the cosy localization database. I J Robotics Res 28(5):588–594

Pronobis A, Fornoni M, Christensen HI, Caputo B (2010a) The Robot Vision Track at ImageCLEF 2010. In: (Braschler et al, 2010)

Pronobis A, Xing L, Caputo B (2010b) Overview of the CLEF 2009 Robot Vision Track. In: (Peters et al, 2010), pp 110–119

Sahbi H (2013) CNRS - TELECOM paristech at imageclef 2013 scalable concept image annotation task: Winning annotations with context dependent svms. In: (Forner et al, 2013)

Sahbi H (2016) CNRS TELECOM paristech at imageclef 2016 scalable concept image annotation task: Overcoming the scarcity of training data. In: (Balog et al, 2016)

Sánchez-Oro J, Montalvo S, Montemayor AS, Pantrigo JJ, Duarte A, Fresno V, Martínez-Unanue R (2013) Urjc&uned at imageclef 2013 photo annotation task. In: (Forner et al, 2013)

Saurer O, Fraundorfer F, Pollefeys M (2010) Visual localization using global visual features and vanishing points. In: (Agosti et al, 2010)

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556, 1409.1556

Tang TH, Fu1 MH, Huang HH, Chen KT, Chen HH (2018) NTU NLP-Lab at ImageCLEFlifelog 2018: Visual Concept Selection with Textual Knowledge for Understanding Activities of Daily Living and Life Moment Retrieval. In: (Cappellato et al, 2018)

Thomee B, Popescu A (2012) Overview of the ImageCLEF 2012 Flickr Photo Annotation and Retrieval Task. In: (Forner et al, 2012)

Tran MT, Truong TD, Dinh-Duy T, Vo-Ho VK, Luong QA, Nguyen VT (2018) Lifelog Moment Retrieval with Visual Concept Fusion and Text-based Query Expansion. In: (Cappellato et al, 2018)

Uijlings JRR, van de Sande KEA, Gevers T, Smeulders AWM (2013) Selective search for object recognition. International Journal of Computer Vision 104(2):154–171

Ünay D, Çataltepe Z, Aksoy S (eds) (2010) Recognizing Patterns in Signals, Speech, Images and Videos - ICPR 2010 Contests, Istanbul, Turkey, August 23-26, 2010, Contest Reports, Lecture Notes in Computer Science, vol 6388, Springer, DOI 10.1007/978-3-642-17711-8, URL https://doi.org/10.1007/978-3-642-17711-8

Ushiku Y, Harada T, Kuniyoshi Y (2012a) Efficient image annotation for automatic sentence generation. In: Babaguchi N, Aizawa K, Smith JR, Satoh S, Plagemann T, Hua X, Yan R (eds) Proceedings of the 20th ACM Multimedia Conference, MM '12, Nara, Japan, October 29 - November 02, 2012, ACM, pp 549–558

Ushiku Y, Muraoka H, Inaba S, Fujisawa T, Yasumoto K, Gunji N, Higuchi T, Hara Y, Harada T, Kuniyoshi Y (2012b) ISI at imageclef 2012: Scalable system for image annotation. In: (Forner et al, 2012)

Vanegas JA, Ovalle JEA, Montenegro JSO, Páez F, Pérez-Rubiano SA, González FA (2014) Mindlab at imageclef 2014: Scalable concept image annotation. In: (Cappellato et al, 2014)

Villegas M, Paredes R (2012) Overview of the ImageCLEF 2012 Scalable Web Image Annotation Task. In: (Forner et al, 2012)

Villegas M, Paredes R (2014) Overview of the ImageCLEF 2014 Scalable Concept Image Annotation Task. In: (Cappellato et al, 2014)

Villegas M, Paredes R, Thomee B (2013) Overview of the ImageCLEF 2013 Scalable Concept Image Annotation Subtask. In: (Forner et al, 2013)

Xing L, Pronobis A (2009) Multi-cue discriminative place recognition. In: (Peters et al, 2010), pp 315–323

Xu X, Shimada A, Taniguchi R (2014) MLIA at imageclfe 2014 scalable concept image annotation challenge. In: (Cappellato et al, 2014)

Zellhöfer D (2012) Overview of the Personal Photo Retrieval Pilot Task at ImageCLEF 2012. In: (Forner et al, 2012)

Zellhöfer D (2013) Overview of the ImageCLEF 2013 Personal Photo Retrieval Subtask. In: (Forner et al, 2013)

Zhou L, Piras L, Riegler M, Boato G, Dang-Nguyen DT, Gurrin C (2017) Organizer Team at ImageCLEFlifelog 2017: Baseline Approaches for Lifelog Retrieval and Summarization. In: (Cappellato et al, 2017)

Zhou L, Piras L, Riegler M, Lux M, Dang-Nguyen1 DT, Gurrin C (2018) An Interactive Lifelog Retrieval System for Activities of Daily Living Understanding. In: (Cappellato et al, 2018)