ImageCLEF 2022: Multimedia Retrieval in Medical, Nature, Fusion, and Internet Applications

Alba G. Seco de Herrera[1], Bogdan Ionescu[2], Henning Müller[3], Renaud Péteri[4], Asma Ben Abacha[5], Christoph M. Friedrich[6], Johannes Rückert[6], Louise Bloch[6], Raphael Brüngel[6], Ahmad Idrissi-Yaghir[6], Henning Schäfer[6], Serge Kozlovski[7], Yashin Dicente Cid[8], Vassili Kovalev[7], Jon Chamberlain[1], Adrian Clark[1], Antonio Campello[9], Hugo Schindler[10], Jérôme Deshayes[10], Adrian Popescu[10], Liviu-Daniel Ştefan[2], Mihai Gabriel Constantin[2], and Mihai Dogariu[2]

[1] University of Essex, UK, alba.garcia@essex.ac.uk
[2] University Politehnica of Bucharest, Romania
[3] University of Applied Sciences Western Switzerland (HES-SO), Switzerland
[4] University of La Rochelle, France
[5] National Library of Medicine, USA
[6] University of Applied Sciences and Arts Dortmund (FH Dortmund), Germany
[7] Belarussian Academy of Sciences, Belarus
[8] University of Warwick, UK
[9] Wellcome Trust, UK
[10] CEA LIST, France

Abstract. ImageCLEF is part of the Conference and Labs of the Evaluation Forum (CLEF) since 2003. CLEF 2022 will take place in Bologna, Italy. ImageCLEF is an ongoing evaluation initiative which promotes the evaluation of technologies for annotation, indexing, and retrieval of visual data with the aim of providing information access to large collections of images in various usage scenarios and domains. In its 20th edition, ImageCLEF will have four main tasks: (i) a *Medical* task addressing concept annotation, caption prediction, and tuberculosis detection; (ii) a *Coral* task addressing the annotation and localisation of substrates in coral reef images; (iii) an *Aware* task addressing the prediction of real-life consequences of online photo sharing; and (iv) a new *Fusion* task addressing late fusion techniques based on the expertise of the pool of classifiers. In 2021, over 100 research groups registered at ImageCLEF with 42 groups submitting more than 250 runs. These numbers show that, despite the COVID-19 pandemic, there is strong interest in the evaluation campaign.

Keywords: User awareness, medical image classification, medical image understanding, coral image annotation and classification, fusion, ImageCLEF benchmarking, annotated data

1 Introduction

ImageCLEF is a benchmarking activity on the cross-language annotation and retrieval of images in the Conference and Labs of the Evaluation Forum (CLEF) [19, 20]. The 20th anniversary of ImageCLEF will take place in Bologna, Italy, in September 2022[11]. The main goal of ImageCLEF is to promote research in the fields of multi-lingual and multi-modal information access evaluation. Hence, a set of benchmarking activities was designed to test different aspects of mono and cross-language information retrieval systems [16, 19, 20]. Both ImageCLEF [28] and also the overall CLEF campaign have important scholarly impact [28, 29].

Since 2018, the AIcrowd[12] platform (previously crowdAI) is used to distribute the data and receive the submitted results. The platform provides access to the data beyond the competition and allows having an online leaderboard.

The following sections introduce the four tasks that are planned for 2022, namely: ImageCLEFmedical, ImageCLEFcoral, ImageCLEFaware, and the new ImageCLEFfusion. Figure 1 captures a few images the specificities of some tasks.

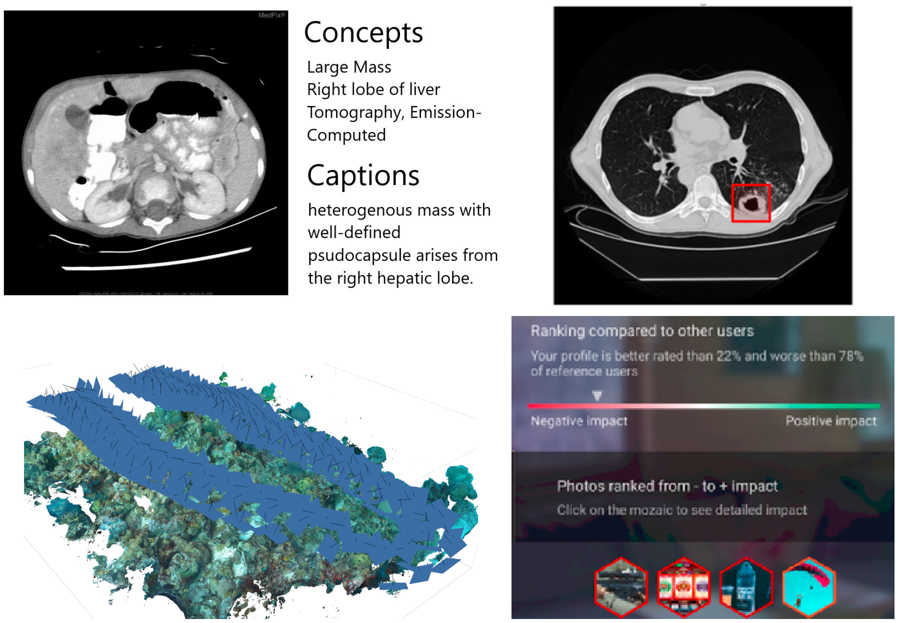
2 ImageCLEFmedical

The ImageCLEFmedical task has been carried out every year since 2004 [20]. The 2022 edition will include two tasks: the caption task, and the tuberculosis task. The *caption* task focuses on interpreting and summarising the insights gained from radiology images. In the 6th edition [12, 13, 22–24] of the task, there will be two subtasks: concept detection and caption prediction. The *concept detection* subtask aims to develop competent systems that are able to predict the Unified Medical Language System (UMLS®) Concept Unique Identifiers (CUIs) based on the visual image content. The F1-Score [15] will be used to evaluate the participating systems in this subtask. The *caption prediction* subtask focuses on implementing models to predict captions for given radiology images. The BLEU [21] score will be used for evaluating this subtask. In 2022, a subset of the Radiology Objects in Context (ROCO) [25] dataset will be used. As in the previous editions, the dataset will be manually curated after using multiple concept extraction methods to retrieve accurate CUIs.

The *tuberculosis* task will be extended from the previous TB-case and/or lesion classification problems [9–11, 17, 18] to the more advanced lesion detection problem. As in previous editions the task will use chest 3D Computed Tomography (CT) scans as source data, but this time participants are expected to detect cavern regions localisation rather than simply provide a label for the CT image. This problem is important because even after successful TB treatment, which satisfies the existing criteria of cavern recovery, patients may still contain colonies of Mycobacterium tuberculosis that could lead to unpredictable disease

[11] https://clef2022.clef-initiative.eu/

[12] https://www.aicrowd.com/


Concepts
Large Mass
Right lobe of liver
Tomography, Emission-Computed
Captions
heterogenous mass with well-defined psudocapsule arises from the right hepatic lobe.

Ranking compared to other users

Your profile is better rated than 22% and worse than 78% of reference users

Negative impact
Positive impact

Photos ranked from - to + impact

Click on the mozaic to see detailed impact

Fig. 1. Sample images from (left to right, top to bottom): ImageCLEFmedical caption with an image with the corresponding CUIs and caption, ImageCLEFmedical tuberculosis with a slice of a chest CT with tuberculosis cavern region, ImageCLEFcoral with 3D reconstruction of a coral reef and ImageCLEFaware with an example of user photos and predicted influence when searching for a bank loan.

relapse. In addition, the subtask of predicting four binary features of caverns suggested by experienced radiologists will be available.

3 ImageCLEFcoral

The increasing use of structure-from-motion photogrammetry for modelling large-scale environments from action cameras attached to drones has driven the next-generation of visualisation techniques that can be used in augmented and virtual reality headsets. Since 2019, the ImageCLEFcoral task addresses the issue automatically annotating these images for monitoring coral reef structure and composition, in support of their conservation. The 4th edition of the task will follow a similar format to previous editions [2–4] where participants automatically segment and label a collection of images that can be used in combination to create three-dimensional models of an underwater environment. Figure 1 shows the 3D reconstruction of a coral reef (approx. 4 × 6 m). To create this model, each image in the subset is represented by a blue rectangle in the image, with the track of multi-camera array clearly visible across the environment. In 2022, this task will contain the same two subtasks as in previous years: *coral reef image annotation and localisation* and *coral reef image pixel-wise parsing*.

The *coral reef image annotation and localisation* subtask requires participants to label the images with types of benthic substrate together with their bounding box. The *coral reef image pixel-wise parsing* subtask requires the participants to segment and parse each coral image into different image regions associated with benthic substrate types. As in previous editions, the performance of the submitted algorithms will be evaluated using the PASCAL VOC style metric of intersection over union (IoU) and the mean of pixel-wise accuracy per class.

Previous editions of ImageCLEFcoral in 2019 and 2020 showed improvements in task performance and promising results on cross-learning between images from different geographical regions. The 3rd edition in 2021 increased the task complexity and size of data available to participants through supplemental data, resulting in lower performance than in previous years. The 4th edition plans to address these issues by targeting algorithms for geographical regions and raising the benchmark performance. As with the 3rd edition, training and test data will form the complete set of images required for 3D reconstruction of the marine environment. This will allow participants to explore novel probabilistic computer vision techniques based on image overlap and transposition of data points.

4 ImageCLEFaware

The online disclosure of personal data often has effects which go beyond the initial context in which data was shared. Content which seems innocuous initially can be interpreted to the users' disadvantage by third parties which have access to the data. For instance, it is now common for prospective employers to search online information about candidates. This process can be done by humans or be based on automatic inferences. Users are entitled to be aware about the potential effects of online data sharing and this task hypothesises that feedback about these effects can be efficiently provided by simulating impactful real-life situations. Since images constitute a large part of the content shared online, the objective of the ImageCLEFaware task is to automatically rate user photographic profiles in four situations in which the users would search for, e.g., a bank loan, an accommodation, a waiter job, or a job in IT.

In the 2nd edition of the task, the dataset will be enriched to include 1000 profiles instead of the 500 included in the 1st edition. Each profile is labelled with an appeal score per situation by several annotators. Participants will be provided with the profiles along with the associated rankings. The objective of the task is to produce an automatic ranking which is as closely correlated as possible to the manual ranking. Correlation will be measured using a classical measure such as the Pearson correlation coefficient.

Task-related resources will be provided by the organisers to encourage participation of different communities. These resources include: (i) visual object ratings per situation obtained through crowdsourcing; (ii) automatically extracted visual object detection for over 350 objects (versus 270 in 2021) which have non-null rating in at least one situation.

In accordance with General Data Protection Regulation, data minimisation is applied and participants receive only the information necessary to carry out the task in an anonymised form. This resources include (i) anonymised visual concept ratings for each situation modelled; (ii) automatically extracted predictions for the images that compose the profiles.

5 ImageCLEFfusion

While the advent of deep learning systems greatly increased the overall performance of computer vision methods in general, there are still some tasks where system performance is low, thus impeding the adoption of automated computer vision methods in the industry. One representative example for this type of tasks is the prediction of subjective properties of multimedia samples. While these tasks may be harder to solve, given the inherent lower annotator agreement present in the datasets associated with such concepts, this task hypothesises that it is possible to significantly improve the current results by using late fusion approaches. *Late fusion*, also knows as *ensembling* or *decision-level fusion*, consists of a set of initial predictors, called *inducers*, that are trained and tested on the dataset, whose prediction outputs are combined in the final step via an *ensembling method* in order to create a new and improved set of predictions. In the current literature late fusion systems are sometimes successfully used even in traditional tasks such as video action recognition [27], and more often in subjective and multimodal tasks like memorability [1], violence detection [7] and media interestingness [30]. Furthermore, latest developments in this field, using deep neural networks as the primary ensembling method show major improvements over traditional ensembling methods, by greatly increasing the performance of individual inducers [5, 6, 26].

In this context, the 1st edition of the ImageCLEFfusion task is proposed. The organisers will provide several task-related resources such as: (i) the datasets that will be used throughout the task; (ii) a large set (more than 15) of pre-computed inducer prediction outputs for the corresponding datasets; and (iii) metrics and data-splits. Participants are tasked with creating novel ensembling methods to significantly increase the performance of the pre-computed inducers. The targeted datasets for the ImageCLEFfusion task will be composed of ground truth data extracted from several subjective multimedia processing tasks like interestingness [8], memorability [14] or result diversification [31].

While the metrics and data-splits will be used to measure ensembling method performance, this task will also look to provide answers to interesting theoretical questions, such as: (i) how does inducer correlation affect ensemble performance; (ii) how does inducer diversity affect ensemble performance; (iii) are there selection methods for inclusion and exclusion of inducers regarding an ensemble; (iv) are deep learning ensembling methods better than other types of approaches? Answering these questions may provide valuable insights for future research, not only with regards to the best performing ensemble methods, but also into the reduction of hardware requirements by inducer selection.

6 Conclusion

This paper presents an overview of the upcoming ImageCLEF at the CLEF 2022. ImageCLEF has been organising many tasks in a variety of domains in the field of visual media analysis, indexing, classification, and retrieval. The 20th anniversary of the task includes a variety of tasks in the fields of medical imaging, nature, system fusion, and internet applications. All the tasks will provide a set of new test collections simulating real-world situations. Such collections are important to enable researchers to assess the performance of their systems and to compare their results with others following a common evaluation framework.

Acknowledgement

Part of this work is supported under the H2020 AI4Media "A European Excellence Centre for Media, Society and Democracy" project, contract #951911. The work of Louise Bloch and Raphael Brüngel was partially funded by a PhD grant from the University of Applied Sciences and Arts Dortmund (FH Dortmund), Germany. The work of Ahmad Idrissi-Yaghir and Henning Schäfer was funded by a PhD grant from the DFG Research Training Group 2535 Knowledge- and data-based personalisation of medicine at the point of care (WisPerMed).

References

1. Azcona, D., Moreu, E., Hu, F., Ward, T.E., Smeaton, A.F.: Predicting media memorability using ensemble models. In: Working Notes Proceedings of the MediaEval 2019 Workshop. CEUR Workshop Proceedings, vol. 2670. CEUR-WS.org (2019)
2. Chamberlain, J., Campello, A., Wright, J.P., Clift, L.G., Clark, A., García Seco de Herrera, A.: Overview of ImageCLEFcoral 2019 task. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2019). CEUR Workshop Proceedings, vol. 2380. CEUR-WS.org (2019)
3. Chamberlain, J., Campello, A., Wright, J.P., Clift, L.G., Clark, A., García Seco de Herrera, A.: Overview of the ImageCLEFcoral 2020 task: Automated coral reef image annotation. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2020). CEUR Workshop Proceedings, vol. 2696. CEUR-WS.org (2020)
4. Chamberlain, J., García Seco de Herrera, A., Campello, A., Clark, A., Oliver, T.A., Moustahfid, H.: Overview of the ImageCLEFcoral 2021 task: Coral reef image annotation of a 3D environment. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2021). CEUR Workshop Proceedings, vol. 2936. CEUR-WS.org (2021)
5. Constantin, M.G., Ştefan, L.D., Ionescu, B.: DeepFusion: Deep ensembles for domain independent system fusion. In: Proceedings of the 27th International Conference on Multimedia Modeling (MMM 2021). vol. 12573, pp. 240–252. LNCS Lecture Notes in Computer Science, Springer (2021)
6. Constantin, M.G., Ştefan, L.D., Ionescu, B., Duong, N.Q., Demarty, C.H., Sjöberg, M.: Visual interestingness prediction: A benchmark framework and literature review. International Journal of Computer Vision pp. 1–25 (2021)

7. Dai, Q., Zhao, R.W., Wu, Z., Wang, X., Gu, Z., Wu, W., Jiang, Y.G.: Fudan-Huawei at MediaEval 2015: Detecting violent scenes and affective impact in movies with deep learning. In: Working Notes Proceedings of the MediaEval 2015 Workshop. CEUR Workshop Proceedings, vol. 1436. CEUR-WS.org (2015)
8. Demarty, C.H., Sjöberg, M., Ionescu, B., Do, T.T., Gygli, M., Duong, N.: MediaEval 2017 predicting media interestingness task. In: Working Notes Proceedings of the MediaEval 2017 Workshop. CEUR Workshop Proceedings, vol. 1984. CEUR-WS.org (2017)
9. Dicente Cid, Y., Kalinovsky, A., Liauchuk, V., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2017 – predicting tuberculosis type and drug resistances. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2017). CEUR Workshop Proceedings, vol. 1866. CEUR-WS.org (2017)
10. Dicente Cid, Y., Liauchuk, V., Klimuk, D., Tarasau, A., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2019 – automatic CT-based report generation and tuberculosis severity assessment. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2019). CEUR Workshop Proceedings, vol. 2380. CEUR-WS.org (2019)
11. Dicente Cid, Y., Liauchuk, V., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2018 – detecting multi-drug resistance, classifying tuberculosis type, and assessing severity score. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2018). CEUR Workshop Proceedings, vol. 2125. CEUR-WS.org (2018)
12. Eickhoff, C., Schwall, I., García Seco de Herrera, A., Müller, H.: Overview of ImageCLEFcaption 2017 – the image caption prediction and concept extraction tasks to understand biomedical images. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2017). CEUR Workshop Proceedings, vol. 1866. CEUR-WS.org (2017)
13. García Seco De Herrera, A., Eickhof, C., Andrearczyk, V., Müller, H.: Overview of the ImageCLEF 2018 caption prediction tasks. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2018). CEUR Workshop Proceedings, vol. 2125. CEUR-WS.org (2018)
14. García Seco De Herrera, A., Kiziltepe, R.S., Chamberlain, J., Constantin, M.G., Demarty, C.H., Doctor, F., Ionescu, B., Smeaton, A.F.: Overview of MediaEval 2020 predicting media memorability task: What makes a video memorable? Working Notes Proceedings of the MediaEval 2020 Workshop (2020)
15. Goutte, C., Gaussier, E.: A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: Advances in Information Retrieval – 27th European Conference on IR Research (ECIR 2005). pp. 345–359. Springer (2005)
16. 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)
17. Kozlovski, S., Liauchuk, V., Dicente Cid, Y., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2021 – CT-based tuberculosis type classification. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2021). CEUR Workshop Proceedings, vol. 2936. CEUR-WS.org (2021)
18. Kozlovski, S., Liauchuk, V., Dicente Cid, Y., Tarasau, A., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2020 – automatic CT-based report generation. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2020). CEUR Workshop Proceedings, vol. 2696. CEUR-WS.org (2020)

19. 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)
20. Müller, H., Kalpathy-Cramer, J., García Seco de Herrera, A.: Experiences from the ImageCLEF medical retrieval and annotation tasks. In: Information Retrieval Evaluation in a Changing World, pp. 231–250. Springer (2019)
21. 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 of the Association for Computational Linguistics (ACL 2002). pp. 311–318 (2002)
22. Pelka, O., Abacha, A.B., García Seco de Herrera, A., Jacutprakart, J., Friedrich, C.M., Müller, H.: Overview of the ImageCLEFmed 2021 concept & caption prediction task. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2021). CEUR Workshop Proceedings, vol. 2936. CEUR-WS.org (2021)
23. Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Müller, H.: Overview of the ImageCLEFmed 2019 concept detection task. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2019). CEUR Workshop Proceedings, vol. 2380. CEUR-WS.org (2019)
24. Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Müller, H.: Overview of the ImageCLEFmed 2020 concept prediction task: Medical image understanding. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2020). CEUR Workshop Proceedings, vol. 2696. CEUR-WS.org (2020)
25. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology Objects in COntext (ROCO): A multimodal image dataset. In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, pp. 180–189. Springer (2018)
26. Ştefan, L.D., Constantin, M.G., Ionescu, B.: System fusion with deep ensembles. In: Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR 2020). pp. 256–260. Association for Computing Machinery (ACM) (2020)
27. Sudhakaran, S., Escalera, S., Lanz, O.: Gate-shift networks for video action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020). pp. 1102–1111 (2020)
28. 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) (2011)
29. 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)
30. Wang, S., Chen, S., Zhao, J., Jin, Q.: Video interestingness prediction based on ranking model. In: Proceedings of the joint workshop of the 4th workshop on affective social multimedia computing and first multi-modal affective computing of large-scale multimedia data (ASMMC-MMAC 2018). pp. 55–61. Association for Computing Machinery (ACM) (2018)
31. Zaharieva, M., Ionescu, B., Gînsca, A.L., Santos, R.L., Müller, H.: Retrieving diverse social images at MediaEval 2017: Challenges, dataset and evaluation. In: Working Notes Proceedings of the MediaEval 2017 Workshop. CEUR Workshop Proceedings, vol. 1984. CEUR-WS.org (2017)