Int J Multimed Info Retr (2016) 5:135–136
DOI 10.1007/s13735-016-0106-7

TRENDS AND SURVEYS

Open and free datasets for multimedia retrieval

Erwin M. Bakker[1]

Received: 21 June 2016 / Accepted: 27 June 2016 / Published online: 15 July 2016

Abstract One of the pillars of the scientific community is open and free datasets. Not only do they allow benchmarking, evaluation and also reproducibility but they also provide an important contribution themselves by allowing researchers to gain deeper insight into the strengths and weaknesses of their algorithms and paradigms. Furthermore, in the cases of the larger datasets, they facilitate major advances (e.g., concept learning using big data). Here, we present some of the recent free and open datasets for the scientific community.

Keywords Open datasets · Multimedia information retrieval · Freely redistributable · Evaluation · Performance

1 Introduction

In the first generation of multimedia retrieval, there were several widely used datasets such as the Corel Stock Photography Collection and the MPEG-7 dataset which became unavailable due to copyrights and other legal hindrances. This caused problems for many new researchers who were told by reviewers that they should use them. We would like to encourage open and freely redistributable datasets so that scientific advances are not hindered by inaccessibility of data.

The datasets mentioned here are limited to ones which are (1) Open—no login or special permission is required for download (this was verified in May 2016, but could change), and (2) Freely Redistributable—the datasets may be redistributed without asking for permission from the authors as long as they are not used for commercial purposes.

✉ Erwin M. Bakker
erwinmbakker@gmail.com; erwin@liacs.nl

[1] LIACS Media Lab, Leiden University, Leiden, The Netherlands

2 Scientific datasets

2.1 Div150Cred: a social image retrieval result diversification with user tagging credibility dataset [1]

With the prevalence of social media platforms, such as Twitter, Facebook and Instagram, searching through social media has become an important area. In social media search, results should be both relevant and diverse and also include social user links. From previous studies it has been found that tourist landmark locations are especially popular with social media. In line with the studies, this dataset comprises 300 landmark locations represented by 45K Flickr photos, 16M photo links by 3000 users with associated metadata and descriptors. Furthermore, it was used in the MediaEval task on retrieving diverse social images.

2.2 Fashion 10000: an enriched social image dataset for fashion and clothing [2]

As multimedia retrieval evolves and matures, new areas are explored. One of these is the fashion and clothing domain. This dataset contains 32 K images including their context and social metadata. Moreover, they also include annotations for various content analysis algorithms based on the Amazon Mechanical Turk. These annotation categories included fashion/clothing related, special clothing item, number of people, professional model, person wearing fashion and formal/informal clothing. In addition, it was used at the Crowdsourcing task in the MediaEval initiative.

2.3 Stanford I2V: a news video dataset for query-by-image experiments [3]

With the growth of video collections, such as NetFlix and Amazon Prime as well as personal mobile devices, one of the major trends is searching for video. An intuitive way of querying video databases is using images—which videos have content similar to the given image? This dataset contains 3800 h of newscast videos from 84 K clips and more than 200 queries with ground truth annotations allowing for larger scale performance evaluation of I2V search algorithms.

2.4 World-wide scale geotagged image dataset for automatic image annotation and reverse geotagging [4]

Geotagged images can give important additional information regarding the image search process. In addition, they can also be used to train the process of automatic location identification of unknown images. This paper presents a large scale geotagged image dataset which contains over 14M Flickr images with the appropriate metadata. The authors designed a crawling strategy ensuring that the data set covers the whole world, and densities reflect the popularity of the locations.

2.5 Common objects in context (MS COCO) [5]

Multimedia retrieval has been in need of precisely annotated imagery for the past decade. The COCO dataset is the largest one so far containing 328 K images with 2.5M labeled instances. It is important to note that their annotations go deeper than simple text keywords. The authors also provide the location of the object or person in the image. In terms of facilitating major advances, this dataset may be the most significant in the coming years.

References

1. Ionescu B, Popescu A, Lupu M, Gînscă AL, Boteanu B, Müller H (2015) Div150Cred: A social image retrieval result diversification with user tagging credibility dataset. In: Proceedings of ACM MMSys'15, Portland, Oregon, pp 207–212. http://skuld.cs.umass.edu/traces/mmsys/2015/paper-5/index.html
2. Loni B, Cheung LY, Riegler M, Bozzon A, Gottlieb L, Larson M (2014) Fashion 10000: An enriched social image dataset for fashion and clothing. In: Proceedings of MMSys'14, Singapore, Singapore, pp 41–46. http://skuld.cs.umass.edu/traces/mmsys/2014/user05.tar
3. Araujo A, Chaves J, Chen D, Angst R, Girod B (2015) Stanford I2V: a news video dataset for query-by-image experiments. In: Proceedings of ACM MMSys'15, Portland, Oregon, pp 237–242. http://purl.stanford.edu/zx935qw7203
4. Mousselly-Sergieh H, Watzinger D, Huber B, Döller M, Egyed-Zsigmond E, Kosch H (2014) World-wide scale geotagged image dataset for automatic image annotation and reverse geotagging. In: Proceedings of MMSys'14, Singapore, Singapore, pp 47–52. http://skuld.cs.umass.edu/traces/mmsys/2014/user03.tar
5. Lin T-Y, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, Perona P, Ramanan D, Lawrence Zitnick C, Dollár P (2015) Microsoft COCO: common objects in context. arxiv:1405.0312. http://mscoco.org/dataset/#download