

## Chapter 2

# The Benchmark as a Research Catalyst: Charting the Progress of Geo-prediction for Social Multimedia

**Martha Larson, Pascal Kelm, Adam Rae, Claudia Hauff, Bart Thomee, Michele Trevisiol, Jaeyoung Choi, Olivier Van Laere, Steven Schockaert, Gareth J.F. Jones, Pavel Serdyukov, Vanessa Murdock and Gerald Friedland**

**Abstract** Benchmarks have the power to bring research communities together to focus on specific research challenges. They drive research forward by making it easier to systematically compare and contrast new solutions, and evaluate their

---

M. Larson (✉) · C. Hauff  
Delft University of Technology, Delft, The Netherlands  
e-mail: m.a.larson@tudelft.nl

C. Hauff  
e-mail: c.hauff@tudelft.nl

P. Kelm  
Technische Universität, Berlin, Germany  
e-mail: kelm@nue.tu-berlin.de

A. Rae  
Future Cities Catapult, London, UK  
e-mail: arae@futurecities.catapult.org.uk

B. Thomee  
Yahoo Labs, San Francisco, CA, USA  
e-mail: bthomee@yahoo-inc.com

M. Trevisiol  
Pompeu Fabra University, Barcelona, Spain  
e-mail: trevisiol@acm.org

O. Van Laere  
Yahoo Labs, Barcelona, Spain  
e-mail: vanlaere@yahoo-inc.com

J. Choi · G. Friedland  
ICSI, Berkeley, CA, USA  
e-mail: jaeyoung@icsi.berkeley.edu

G. Friedland  
e-mail: fractor@icsi.berkeley.edu

S. Schockaert  
Cardiff University, Cardiff, UK  
e-mail: s.schockaert@cs.cardiff.ac.uk

performance with respect to the existing state of the art. In this chapter, we present a retrospective on the Placing Task, a yearly challenge offered by the MediaEval Multimedia Benchmark. The Placing Task, launched in 2010, is a benchmarking task that requires participants to develop algorithms that automatically predict the geolocation of social multimedia (videos and images). This chapter covers the editions of the Placing Task offered in 2010–2013, and also presents an outlook onto 2014. We present the formulation of the task and the task dataset for each year, tracing the design decisions that were made by the organizers, and how each year built on the previous year. Finally, we provide a summary of future directions and challenges for multimodal geolocation, and concluding remarks on how benchmarking has catalyzed research progress in the research area of geolocation prediction for social multimedia.

## 2.1 Introduction

A benchmark is a standardized task that is carried out in order to evaluate alternative approaches to addressing the task and to facilitate a fair comparison between multiple strategies for tackling this task. Benchmarks bring research communities together to focus on a specific research challenge. This coming together of researchers with common interests to work on a specific task can drive research forward by enabling them to comparatively evaluate their work. In this chapter, we present a retrospective of the Placing Task, a challenge offered within the MediaEval Benchmarking Initiative for Multimedia Evaluation.<sup>1</sup> We track the development of the Placing Task over four editions from 2010 to 2013, and present an outlook to 2014. The evolution of the Placing Task in MediaEval illustrates the power of a benchmark to establish a new research topic and a community of collaborating researchers working to address the challenges of this topic, together with persistent datasets that enable researchers to evaluate their results with respect to the state of the art, and explore the effectiveness of the new algorithms that they develop.

---

<sup>1</sup> <http://multimediaeval.org>.

G.J.F. Jones  
Dublin City University, Dublin, Ireland  
e-mail: Gareth.Jones@computing.dcu.ie

P. Serdyukov  
Yandex, Moscow, Russia  
e-mail: pavser@yandex-team.ru

V. Murdock  
Microsoft, Bellevue, WA, USA  
e-mail: vanmur@microsoft.com

Geoprediction for social multimedia, also known as *placing*, is the task of inferring geocoordinates for images or videos that users have uploaded to social sharing websites. The key application of geoprediction technology is indexing images and videos online, making them easier, to find, manage, and browse, and, in general, more useful for users. Location, and concepts related to location, have a close connection with how users interpret, organize, and use multimedia, and consequently applications that use geocoordinates are considered to be important in allowing users to get the most out of social multimedia. Many of today’s cameras and phones can and do record geoinformation. Nonetheless, a large number of videos and images are uploaded without georeference. For this reason, high-performance placing algorithms are necessary in order to generate metadata that makes it easier for users to retrieve and browse social multimedia.

This chapter discusses Placing Task, a challenge offered by the MediaEval Benchmarking Initiative for Multimedia Evaluation<sup>2</sup> to the multimedia research community with the goal of fostering the development of new algorithms addressing the task of automatically predicting the geocoordinates of social multimedia. While this chapter focuses on the topic of geoprediction, we anticipate that is also relevant to multimedia benchmarking and general. We hope that it might ultimately serve to support the consolidation and catalyzation of results in other research areas, which may benefit from applying the strategies and techniques used by the MediaEval Placing Task.

### ***2.1.1 The Placing Challenge for Social Multimedia***

The nature of social multimedia means that placing is fundamentally a multimedia challenge, involving different modalities. Images and videos uploaded by users are associated with metadata, such as titles and descriptions. They are also associated with user-contributed tags and comments. Often the user-uploaded multimedia items are connected in a social network: here, information such as social connections and views may also be available. In the case of video, the multimedia signal involves temporal patterns which can be exploited. Video typically involves both a visual and an audio channel. Audio includes spoken content, but also music and environmental sounds contained in videos.

Interest in technologies that infer geocoordinates was established by work such as [22] and [16]. Research effort devoted specifically to placing multimedia that users share on the Internet gained momentum along with the rise of social media. Geoprediction for social multimedia took on an independent form as a task in its own right, with the publication of a paper entitled, “Placing Flickr Photos on a Map” [52], which followed on the heels of [10]. Due to the influence of [52], the word “Placing” was adopted as the name of the task in the MediaEval benchmark.

It is important to note that the task of placing social multimedia is different from predicting geolocation of multimedia data that was not captured by users for personal use or social sharing. The phenomenon of people taking and sharing pictures

---

<sup>2</sup> <http://multimediaeval.org>.

and videos ranges from people who point-and-shoot in order to capture a moment or memory, to people who document events and objects, and people who pursue photography as a hobby. For whatever reason that people produce and distribute multimedia, it is clear that taking a picture or capturing a video is not a random act. Rather it occurs with a reason. The result is that social multimedia is characterized by a particular distribution of subject matter and of photographic style. In short, multimedia shared on the Internet can be considered a *social signal*, i.e., an information stream whose characteristics are determined by the underlying behavior of the people who produce it.

The challenge of placing social multimedia has multiple dimensions: First, placing algorithms must be able to confront the noise and uncertainty associated with social multimedia. The relationship between the visual content of an image and the location at which the image was taken is often a weak one. For example, two images can both depict a Black Labrador in front of a red Volkswagen. The content of both images is visually distinctive, making the images potentially very similar to each other with respect to image processing algorithms. Yet, it is possible that the two were taken many kilometers apart. Conversely, two images taken at the same location, for example, a panorama and a close-up shot, may have no visual content in common. Another source of uncertainty is user-contributed metadata: titles, descriptions, and tags often receive only little attention from users uploading photos and can be incomplete or completely misleading.

Second, algorithms must be capable of effectively combining multiple modalities. Noise and uncertainty can be addressed by simultaneously exploiting multiple information sources. However, in order to benefit from the availability of multiple modalities, multimedia placing algorithms must be able to effectively exploit the complementary information that they contain. The contribution of each modality should enhance the ability of the algorithm to distinguish location. In the extreme cases, such exploitation requires the ability to identify cases in which one modality should be trusted and the others ignored.

Third, placing algorithms must be able to exploit large quantities of data. Geoprediction for social multimedia requires building algorithms that can place a multimedia item at any point in the world. The ability of algorithms to predict place reliably rests on their capacity to process and exploit the large amounts of social multimedia data that are available online, if they are to maximize their ability to cover the world's surface.

Finally, placing algorithms must be able to deal with the uneven distribution of the geotagged data that is available for training. Many locations are associated with rich resources in the form of a large number of multimedia items that have been taken there, and are associated with geocoordinates and available online. Other locations are represented by little to no data, creating an overall data sparseness problem. Fully facing the challenge of "placing" requires developing algorithms that explore a range of different techniques so that it is possible to cover each of these dimensions.

In this chapter, we trace the rise of interest and attention in the multimedia community to the task of placing multimedia items on the world map. Placing has received attention from a broad spectrum of researchers. However, a unique community of

researchers has emerged who have worked together to actively define the task of placing social multimedia and to guide and pursue the development of placing solutions. This community, of which key members are the authors of this chapter, have used benchmarking as a tool to encourage and guide the development of algorithms that address the task of multimodal geolocation estimation for social media. Specifically, they have launched and grown the benchmarking task “Placing: Geocoordinate Prediction for Social Multimedia” within the MediaEval Benchmarking Initiative for Multimedia Evaluation.<sup>3</sup> In this chapter, we discuss the major developments in algorithms to address the challenges of “placing,” and how these developments have been guided by the Placing Task at the MediaEval benchmark.

This chapter charts the development of the Placing Task over the years 2010–2014. In Sect. 2.1.2, we briefly discuss the history of benchmarking, and its ability to promote progress in specific areas of research. Then, in Sect. 2.2, we discuss the design of the task in each year and mentions major results. Based on the experiences in the benchmark, Sect. 2.3 presents an overview for the future challenges that are faced by researchers in the area of Placing. Finally, Sect. 2.4 finishes with a conclusion and outlook.

### ***2.1.2 The Benefits of Benchmarking***

A benchmark is a standardized task that is carried out in order to make possible a fair comparison among multiple algorithms. A benchmark generally consists of a description of a task, resources needed to address that task, and a standard metric or evaluation procedure used to judge the quality of algorithms that address the task. Although other areas have a slightly different definition, this definition holds for the fields of information retrieval and multimedia, from which the Placing Task draws most of its participants.

The roots of the benchmark evaluation movement that gave rise to the MediaEval Multimedia Benchmark can be traced to TREC, the Text REtrieval Conference,<sup>4</sup> which was established in 1992 by the US National Institute of Standards and Technology (NIST). TREC focused initially on text retrieval tasks, but has progressed to topics such as web search, various social media search tasks, and speech and video search, the latter within the TRECVideo benchmark. TREC was followed by the establishment in 1999 of NTCIR<sup>5</sup> in East Asia and in 2000 of CLEF—Cross Language Evaluation Forum in Europe.<sup>6</sup> MediaEval was founded in 2008 as a track named “VideoCLEF” within CLEF. In 2008–2009, VideoCLEF ran tasks examining automatic video tagging and cross-language video search. In 2010, VideoCLEF

---

<sup>3</sup> <http://www.multimediaeval.org>.

<sup>4</sup> <http://trec.nist.gov>.

<sup>5</sup> <http://research.nii.ac.jp/ntcir/>.

<sup>6</sup> <http://www.clef-campaign.org>.

expanded its task offering and became an independent benchmarking initiative named MediaEval. In this year, the Placing Task was established within MediaEval.

The most often cited benefits of benchmarks is that they concentrate research effort on specific problems or challenges, enable cross-site comparison of approaches to address these problems, and drive forward the state of the art with respect to understanding methods or development of new ones. These benefits have been discussed specifically in relation to MediaEval in [36] and [34]. Another important benefit of benchmarks is that they provide stakeholder, e.g., companies who are interested in developing products based on placing technology, with a way to overview and to compare different solutions. Involving stakeholders in benchmarking can couple technologies to real-world needs [45].

One of the less-discussed benefits of benchmarking is that it allows researchers to learn how to deal with a new type of problem. Participating in a benchmark enables researchers to expand the horizons or scope of their research. The opportunities presented by benchmarks are valuable for both students who are developing thesis topics, and also for experienced researchers in the field who are interested in broadening the scope of their investigations. As a multimedia benchmark, MediaEval is related to TRECVideo [54]. TRECVideo is a video retrieval benchmark was first established as TREC track 2001–2002 and became an independent benchmark in 2003. Traditionally, TRECVideo has been more closely tied to the signal processing aspects of multimedia challenges. MediaEval distinguishes itself from TRECVideo by focusing on the human and social aspects of multimedia.

MediaEval follows a yearly cycle, like many other benchmarks. However, it is distinct in important aspects, which we discuss briefly here.

First, the tasks that are offered by MediaEval are decided on the basis of a community survey. This survey gives details of the tasks potentially available for the next round of the benchmark, and is circulated widely at the beginning of the year. The survey gives details of the offered task, but frequently also different possibilities for test data might be used or the specific research questions to be addressed. The survey gathers information about which tasks researchers and stakeholders find most interesting, and which aspects of the tasks that would be interested in focusing on. In the case of the Placing Task, the survey has allowed us to determine that researchers appreciate extra resources, such as visual features, to be released along with the task data, and has also made it possible to ensure that the dataset is structured so that it is computationally feasible for researchers in a wide variety of research labs to tackle the task.

Second, MediaEval tasks are largely autonomous from the overall benchmark, whose role is limited to ensuring the smooth running of the overall activity by maintaining the schedule for the year, and organizing the yearly workshop. This autonomy means that the entire responsibility for developing the task description and dataset, organizing submission of the “runs” (i.e., the results of experimental conditions) submitted by the task participants and their evaluation lies in the hands of the task organizers. In this way, the task organizers have the freedom to develop the most productive and useful task possible, but they also bear the responsibility of ensuring that the dataset is released on time, and that the task is successful.

Third, the MediaEval workshop is seen as a gathering place at which teams that have developed solutions to the tasks come together to exchange ideas and experiences and to forge communities with shared research interests. MediaEval encourages task participants to make use of the resources developed for the task and participants; shared experiences in tackling the task to move beyond the results reported in the Working Notes proceedings published at the workshop. After the workshop task participants, either as individual teams or in combined groups publish papers at mainstream venues. Placing is one example of a new task which has developed an active ongoing research community of shared interested through its participation in the MediaEval benchmark. The communities are typically highly energetic and are comprised of both established researchers and those developing their careers, such as postdoctoral researchers and those beginning their research journey, such as PhD students. Publication of MediaEval results outside the MediaEval workshop promotes the work of MediaEval to wider research communities.

A benchmark is useful if it allows researchers to make progress on a specific problem. It is even more useful if it allows researchers to make gain insight into a whole new class of problems. We believe that multimedia placing represents a new class of multimedia problems. Here, we briefly mention the aspects of the Placing Task that have led us to the conclusion that developing approaches to address the Placing Task will contribute to moving the field of multimedia forward as a whole.

Placing a photo on the world map involves inferring meaning from huge quantities of multimedia data, qualifying it as the type of problem currently often referred to be the term *Big Data*. As mentioned above, it is a fundamentally multimodal problem: creating solutions to address the Placing Task requires developing effective approaches for combining modalities. The data has been created by humans, and, as such, is characterized by an underlying social signal that reflects human behavior, such as travel and photo taking habits. The data used in the Placing Task is “found data.” In other words, it is not collected under controlled conditions, but rather gathered in the wild. Finally, Placing Task algorithms build successfully on “classic” machine learning and information retrieval algorithms. For this reason, results of the Placing Task are likely to find application in other areas of the fields of multimedia and information retrieval.

## 2.2 Charting the Progress

In this section, we discuss, in turn, each year that the Placing Task has been offered at the MediaEval benchmark. The discussion documents the process of building each new year of the benchmark on the following year. Each year chose a different direction in which to push beyond the challenge that was addressed by the previous year into potentially productive, however, yet uncharted territory.

### 2.2.1 *Placing Task 2010: Inception*

The inception of the MediaEval Placing Task dates to 2010, the year that MediaEval became an independent benchmark. The task was launched by Pavel Serdyukov and Vanessa Murdock in the wake of their 2009 SIGIR paper entitled “Placing Flickr Photos on a Map” [52]. The paper, and its follow-up [44], showed the strength of methods that exploit metadata assigned by users to images in order to predict geocoordinates.

Previously, Flickr had started to allow users to upload videos of up to 90 s in length in addition to photos. The tagline associated by Flickr with the introduction of video was, “It is like a photo, but it moves!” Both the size limit and the tagline suggest that it should not be assumed a priori that Flickr videos have the same characteristics as videos uploaded elsewhere on the Internet. Instead, Flickr videos could be expected to have much in common with photographs.

The original paper had investigated photos, and the task was designed to investigate the problem of placing Flickr videos on the map. Use of open resources such as gazetteers or Wikipedia were encouraged. This made it possible to compare approaches that try to extract place names from user-contributed metadata using gazetteers to approaches that build statistical models of text features. Also, use of images to help predict the geolocation of videos was encouraged. In this sense, the task could be considered to be related to computer vision work carried out on scene matching. The ICCV 2005 Computer Vision Contest was called “Where am I?”<sup>7</sup> and required participants to use visual features to match images with unknown locations to images whose locations were known. In contrast to the Placing Task, this contest tackled a small-scale problem (in the order of hundreds of images), and also did not use social images shared online.

#### 2.2.1.1 **Design and Data**

The MediaEval 2010 Placing Task<sup>8</sup> was carried out using a dataset of Creative Commons licensed videos crawled from Flickr containing 5125 videos in the training set and 5091 videos in the test set. The metadata for each video included user-contributed title, tags, description comments, and also information about the user who uploaded the video (including favorites and contacts). The metadata for ~3 Million Creative Commons licensed, geotagged Flickr images was also included. Participants wishing to use visual features could use this metadata to download the images from Flickr.

A set of basic visual features extracted for all images and for keyframes of the videos was provided. Such a resource makes it possible for teams to participate in the benchmark that do not have the infrastructure necessary for large-scale visual feature extraction. Additionally, it ensures that everyone has access to features that were extracted using the same implementation of the feature extractor. In this way,

---

<sup>7</sup> <http://research.microsoft.com/en-us/um/people/szeliski/visioncontest05/default.htm>.

<sup>8</sup> <http://multimediaeval.org/mediaeval2010/placing>.

it is possible to control for the impact of minor variations in feature extraction on the comparison of geolocation approaches.

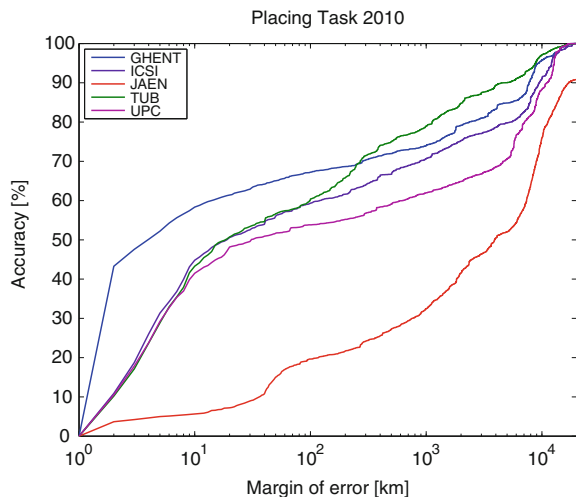
There are 16 zoom levels with which users can place multimedia items on the map in Flickr, and this corresponds to 16 accuracy levels between 1 (world-level) and 16 (street level). The videos in the dataset were selected to provide a broad coverage of users, but also because they have a high accuracy at the street level. The location at which the uploading user had placed the videos on the map was used as the ground truth. Performance was measured in terms of the distance between the location hypothesized by the experimental system and the reference location, i.e., ground truth. The results were reported in terms of the number of photos correctly placed within a 1, 5, 10, 50 and 100km radius.

### 2.2.1.2 Results and Insights

Five groups “crossed the finish line” for this task in MediaEval 2010, submitting runs and also a working notes paper. Their papers can be found in the MediaEval 2010 Working Notes Proceedings [33]. The best results for each team are plotted in Fig. 2.1. Note that for the maximum evaluation radius, an approach will achieve 100% accuracy unless it fails to produce a prediction for all multimedia items.

Here we summarize the major results in 2010. The best performing run was submitted by Ghent University, which used a two-step approach that made use of the textual metadata associated with the images. In the first step, a language model identified the most likely area of the video. In the second step, the location of the video was pinpointed by identifying the closest resources (images and videos) from the training set. The results were reported in the working notes paper [58] and also in [59].

**Fig. 2.1** Participants of the Placing Task 2010



International Computer Science Institute (ICSI) proposed two approaches [7]. The first exploited the prior distribution of tags, in particular choosing tag candidates extracted from video metadata on the basis of small spatial variance. The second approach also undertook supervised resolution of toponyms (using Geonames<sup>9</sup>).

The Universitat Politècnica de Catalunya (UPC) pursued an approach applying knowledge resources (a placenames database) combined with natural language processing used to detect and disambiguate geographical names in the metadata [12]. The SINAI group from the University of Jaen applied geographic-named entity recognizer [23].

The only group to make use of visual features in 2010 was the Technische Universität Berlin, who took a grid-based approach to the task, predicting the grid position of the video by combining a textual model based on metadata and a visual model based on low-level visual features [24, 26]. It is interesting to observe that this initial effort by a single group proved “contagious” and in future years more groups moved on to attempt to exploit not only visual but also audio features.

In the first year of the Placing Task, the importance of the user signal became clear. If a user has taken multiple images in the same place, and tagged them with the same tag set, the presence of one of these images in the training data will be enough to correctly identify the location of any of the other images in the test set. The challenge of placing reduces to a matching problem. Flickr allows users to carry out a so-called “bulk upload,” i.e., uploading multiple items at the same time and assigning them the same tag set. On the one hand, “bulk uploads” can be seen as part of the overall problem of prediction the geolocation of multimedia items on Flickr. On the other hand, the fact that bulk uploads occur is strongly influenced by the technical possibilities offered by the platform. For this reason, the patterns in the data caused by bulk uploads cannot be directly attributed to the underlying social signal, i.e., human photo-taking behavior in the real world. Further, participants discovered that if the user specified a home location, this location provided a good fallback location. Systems placed a photo at the home location if there was no other way to predict location.

Already in the first year, the Placing Task embraced the model of publishing initial results in the MediaEval working note proceedings, but not encouraging the working notes paper to be a final product. Instead, researchers attended the benchmark, discussed the results, and wrote revised papers, which they submitted to other venues. The task contributed to a session called “Automatic Tagging and Geotagging in Video Collections and Communities” [34]. The conference included a paper from the Ghent University team concerning their two-step metadata-based approach [59]. It also included a paper from the Technische Universität Berlin team on their combined textual/visual approach [26]. Other publications were directed at ACM Multimedia 2011 and its satellite workshops [15, 25] and IEEE conferences [5, 37]. The presentation and discussion of the results at these conferences contributed to the growth of the task at future years of MediaEval.

---

<sup>9</sup> <http://www.geonames.org>.

An important insight of the first year was that participants were quickly tempted to download the dataset and start to implement approaches without reading the related work. The task organizers were concerned that participants' efforts would then result in them "re-inventing the wheel". In future years, the task description was accompanied by a list of papers entitled "Recommended Reading" and often even referred to as "Required Reading" by the task organizers. This practice was surprisingly helpful, and allowed teams in future years to boost the level of sophistication with which they approached the task.

## ***2.2.2 Placing Task 2011: Consolidation***

The main goal of MediaEval 2011 Placing Task<sup>10</sup> was to consolidate the task of geolocation prediction for social video by building on the 2010 edition. Pascal Kelm and Adam Rae joined Vanessa Murdock and Pavel Serdyukov as organizers of the task.

### **2.2.2.1 Design and Data**

The videos in the test and development sets of 2010 were combined to create the 2011 development set (total of 10,216 videos). Again, metadata for ~3 Million Creative Commons licensed Flickr images was also released for use in training, and task participants could use the metadata or also crawl images from this set from Flickr to develop their systems. Also, a basic set of visual features was again released. The test data consisted exclusively of video and contained 5,347 individual videos. As in 2010, the results were reported as the number of photos correctly placed within a 1, 5, 10, 50, and 100km radius. Additional details can be found in [50].

### **2.2.2.2 Results and Insights**

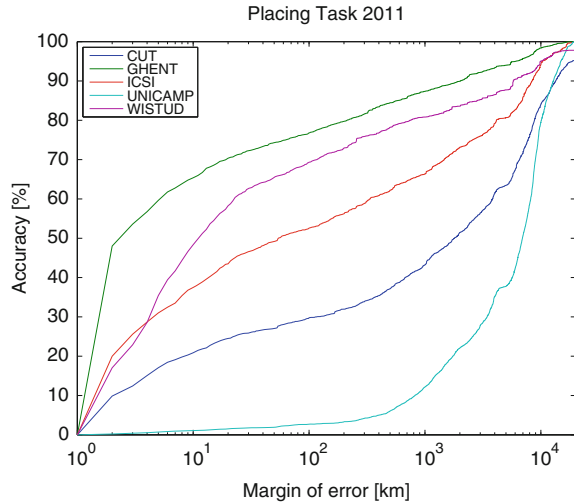
Six groups submitted results for the Placing Task 2011, and their papers can be found in the MediaEval 2011 Working Notes Proceedings [32]. ICSI, Ghent, and UPC returned to participate in the task again, and were joined by University of Campinas (UNICAMP), Delft University of Technology (WISTUD), and Chemnitz University of Technology (CUT). The best scoring runs of each of these groups is plotted in Fig. 2.2.

UPC [13] focused on textual features, an improvement of their 2010 approach [12], and also extended to explore and information retrieval approach. Their best scoring approach exploited a combination of the two. Chemnitz University of Technology [29] applied a text-based retrieval approach.

---

<sup>10</sup> <http://multimediaeval.org/mediaeval2011/placing2011/>.

**Fig. 2.2** Participants of the Placing Task 2011



ICSI proposed an approach integrating textual tags with visual and audio cues [9]. Their best scoring approach used tags refined with visual features.

WISTUD [18] used both textual and visual features, but achieved its best performance using filtered terms. Textual terms that were used by few users or that had a very high geographical spread were filtered out. Interestingly, at the 1 km scale, filtering was not useful.

The innovation of UNICAMP [39] was to exploit the visual content of the video. They used a Histogram of Motion approach [1]. Although the approach could not compete with text-only approaches, it was an important contribution because it represented an effort to use the temporal patterns in video in order to tackle the task. Other groups using visual features had confined themselves to static features derived from individual keyframes.

Ghent [60] extended their textual-metadata approach of 2010 [58]. They successfully exploited additional resources gathered from Flickr (mutually exclusive with the test set). Further, they built on an insight gained in 2010, namely that the granularity of the georegions used by the algorithm is critical. They use Dempster–Shafer theory to determine if there is sufficient evidence for an image to be placed using a given level of granularity [61]. The user’s home location and visual similarity were used as fall back features.

As in the previous year, after the workshop, several of the participants improved and consolidated their approaches and published results in mainstream venue. Papers included the work of Claudia Hauff at TU Delft on exploiting information extracted from microblog posts (i.e., Twitter) and geographical priors [19, 20]. Also, ICSI developed a graph-based approach aimed at dealing with data sparsity [6], and UniCamp extended their approach [46]. Olivier Van Laere and colleagues from the

University of Ghent published a systematic overview of their approaches to the MediaEval 2010 and 2011 Placing Tasks as [62].

In 2011, for the first time, the task was described with an overview paper [50]. There were two reasons for the introduction of overview papers in MediaEval 2011. First, the overview paper describes the task, and communicates to the participants what the organizers are trying to achieve with the task. Second, participants were asked to focus their working notes paper on the unique contribution of their algorithm, and on analyzing their results. They can cite the overview paper for the formulation of the task and a description of the dataset.

In 2011, the Placing Task became interested in the impact of the identity of the uploader on the performance of algorithms. It was observed that the items uploaded by the same person are often near duplicates and/or have the same tag sets. For this reason, it is generally easier to predict the location of an item uploaded by a user, if other items uploaded by that user are included in the training set.

Note that in any given social multimedia collection, it can be expected that users upload multiple items and that some users upload substantially more than others. As a whole, user uploading patterns are part of the underlying characteristics of the social multimedia collection under study. However, the amount of user overlap between the test and the training set is not a fundamental property of the collection. Rather, it has strong dependencies on the data collection process. First, it is dependent on the amount of time that has elapsed between when the test set and when the training set is crawled. Second, it is dependent on the specific characteristics of the users that upload multimedia items in high volumes. Specific characteristics of one or two of these users (e.g., one of them happen to only upload multimedia captured in London), can have a disproportionately large impact on geolocation prediction results.

Although in 2011, the organizers became aware of the impact of exploiting “same-uploader” items for geolocation prediction, it was not until 2013 that the Placing Task issued a dataset in which the training and the test data were uploaded by mutually exclusive sets of users.

### ***2.2.3 Placing Task 2012: Expansion***

The MediaEval 2012 Placing Task<sup>11</sup> again addressed the challenge of generating geolocation predictions for video. It focused on extending the dataset from previous years. In order to encourage participants to develop methods exploiting video content, the task organizers required teams to submit one algorithm that used only video/audio features.

In 2012, Pascal Kelm and Adam Rae served as task organizers. Within the larger context of MediaEval, 2012 saw the handover of both the Placing Task and the Tagging Task (dedicated to automatically generating topical tags for video [51]) to organizer groups that did not include the original founders of the task, but rather were

---

<sup>11</sup> <http://multimediaeval.org/mediaeval2012/placing2012>.



The graph-based approach of ICSI [4] carried out a joint estimation of all the locations of the test videos. Joint estimation was shown to lead to a performance improvement. Another notable contribution was the attempt of CEA LIST [48] to exploit motion features extracted from the videos.

Other approaches to the 2012 Placing Task combined textual and visual [40, 56, 63]. The Ghent University submission [63] built on its system from the previous year, exploring new feature selection and similarity search, but also experimenting with visual features. The UNICAMP approach is notable because it aggregated ranked-lists derived from various modalities. It was extended and subsequently published as [38]. The INRIA/IRISA approach was extended into an approach published as [57]. This approach exploits two-stage hypothesis refinement, which has proven effective in a number of different variants proposed to address the MediaEval Placing Task. It also exploits user’s upload history, social network, and a visual-based matching technique, as well as visual similarity.

TUD proposed a visual-only approach that attempted, with unfortunately little success, to exploit geographical information [42]. Underlying geographical regions related to climate and anthropogenic biomes was used to create regions of the world, on the basis of the assumption that such regions would prove to be visually stable and lend themselves well to modeling.

TU Berlin [27] participated in the MediaEval 2013 Placing Task as an organizer team. As is customary practice within MediaEval, organizers do not report their results in the overall ranking.

Placing Task 2012 was the first time that participants shared code. Because this was also a first for MediaEval, Adam Rae was invited to give a talk at the MediaEval workshop on the importance of code sharing practices. He entitled his talk, “MediaEval Code of Conduct”. The talk was aimed at reinforcing with the community the importance of releasing code.

### ***2.2.4 Placing Task 2013: Volume***

The organization of the MediaEval 2013 Placing Task<sup>12</sup> was taken over by a new team, Claudia Hauff, Bart Thomee, and Michele Trevisiol. The team made four crucial changes to the dataset, as well as the overall task compared to previous years [21]: (i) The task switched from predicting the geolocation of video to predicting the geolocation of images (a return to the original task tackled by [52]); (ii) in the 2013 edition, data collection was carried out in a way that ensured that the set of users contributing images to the training set was mutually exclusive with the set of users contributing images to the test set; (iii) the test set size was several orders of magnitude larger than in previous years to offer new *computational* as well as *algorithmic* challenges; and,

---

<sup>12</sup> <http://multimediaeval.org/mediaeval2013/placing2013>.

(iv) the new subtask of *Placeability* asks task participants to estimate the accuracy of the predicted locations, a necessary metric when employing item location prediction as a preprocessing step for an application.

### 2.2.4.1 Design and Data

*Main task.* Due to the changes described above, the existing datasets could not be reused. The 2013 PlacingTask dataset<sup>13</sup> contains nearly nine million Flickr images released by the owners with Creative Commons License. The sampling process ensured that regions which are popular by Flickr users are also represented as such in the dataset. The extracted metadata and extracted low-level image features follow the templates provided in previous editions, as does the evaluation.

The 2013 Placing Task pioneered the *Russian Dolls* approach to test set creation. Under this approach, the test data is available in five different sizes, in order to allow participation in the task even without very powerful computational resources—ranging from 5,300 images in the smallest test set and 262,000 images in the largest test set. The Russian Dolls approach ensures that the images of a smaller test set are also available in the larger test set as well.

*Placeability subtask.* The Placeability subtask was introduced to investigate whether it is possible to derive a measure of confidence for a predicted location. Past years' experiences had shown that many items can be placed with high accuracy; in 2013, the task made the next step and investigate whether we can also place items with high confidence. Based on the estimated confidence, in self-training, for instance, we may enlarge the training data only with those items that have been located accurately with high confidence, and in the location estimation task we can direct our computational resources to those items, whose confidence score is low.

For a first evaluation, task participants are asked to estimate the error for each prediction in kilometers. The linear and rank correlation coefficients are employed to compare the ability of the algorithms to estimate the error correctly: the true error distance in kilometers (as determined for the main task) is correlated with the predicted error distance. A high correlation coefficient indicates that the algorithm is able to infer the accuracy of the estimation.

As an example, a basic placeability approach can be the following: let us assume that the location estimation approach computes a ranked list of locations (and the top location is returned as estimated location). If the top  $n$  locations are distributed all over the globe, the method may have low confidence and thus the estimated error would be high. On the other hand, if the top  $n$  locations are spatially very close (e.g., having a standard deviation of a few kilometers), then the method's confidence in the location estimate would be high and the error thus low. To derive an error estimate in kilometers the mean distance among the top  $n$  ranked locations can be employed.

---

<sup>13</sup> Available for download: <http://www.st.ewi.tudelft.nl/~hauff/placingTask2013Data.html>.

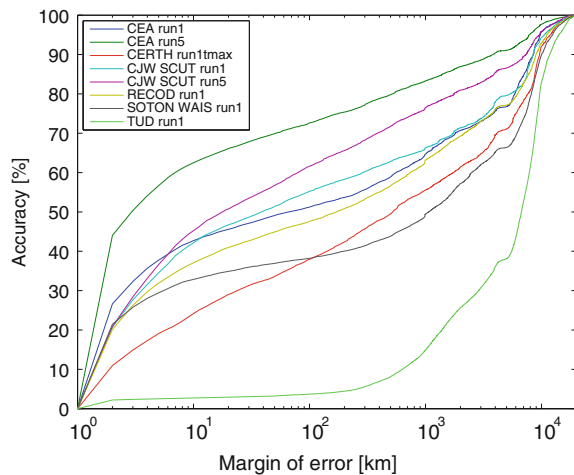
### 2.2.4.2 Results and Insights

An overview of all Placing Task 2013 submissions, including details concerning the specific runs discussed in this section, can be found in the proceedings [31]. In total, 30 runs were submitted by seven different participants. Additionally, the organizers provided two baseline runs.<sup>14</sup> While four participants [11, 28, 43, 47] were able to process the largest test set, two participants [3, 41] opted for the middle ground (processing the third-largest test set with 53,000 test images) and one participant [55] processed the smallest test set. Of the submitted runs, 12 relied on textual metadata, 8 exploited a combination of textual and visual information, and 10 runs used visual information as sole features for location prediction.

As in previous years, relying solely on visual features was not competitive with text features. The best performing submission [43] in this category was able to correctly place less than 5% of the test images (across all test set sizes) within 100km of the true location, the maximum error distance we consider useful for the purpose of employing location prediction in practical settings. This approach employed a two-step process that first retrieved candidate images that were visually similar to the target image, and then refined the geoprediction hypothesis by considering additional visually similar images taken in the neighborhood of the candidate images.

In Fig. 2.4 we show the results of a number of submissions on the test set containing 53,000 images. Several points can be made about this figure. First, we note that CEA\_run5 [47] not only utilizes the provided dataset, but a large amount of additional training data was crawled (90 million items) as well. Most importantly, this includes images from users that are present in our test set, which means that this run exploits the same-user signal.

**Fig. 2.4** Overview of a selected number of Placing Task 2013 submissions (evaluated on the test set containing 53,000 images)



<sup>14</sup> The baseline runs used out-of-the-box location prediction software: <https://github.com/chauff/ImageLocationEstimation>, with geographic filtering enabled.

Further, we point out that CJW\_SCUT\_run5 [3], although not relying on external training data, makes use of two aspects: the user’s home location (if available) as well as the relationship between the test images themselves. A test user might contribute 100 images to the test set with only 20 having any tags associated with them—under this approach, the locations predicted for images with tags are also distributed to those images without tags if they have been taken in a short time frame. While the home location has been employed in previous years, the use of the relationship of images *within* the test set is relatively new, having been previously exploited by ICSI in 2012 [4].

Finally, run CJW\_SCUT\_run5 [3] shows that large gains are also possible when considering the relationship between test images themselves as well as additional user information. In contrast, most participants in 2013 who exploited textual metadata focused on the best way to model regions, although the change in effectiveness across the different modeling strategies was small.

The MediaEval 2013 Placing Task yielded several important overall insights. First, the organizers’ baseline (based on an algorithm proposed in 2011) performed very well: across all runs and evaluation metrics relying solely on the provided data set, it always ranked second or third. This is a somewhat disappointing result, and indicates that forward progress in the task is quite slow.

Second, the test set, although randomly sampled, has a strong influence on the reported metrics, especially for the *ErrorMedian* metric, which is less stable than all metrics relying on a radius based error measure. As an example, while the median error of `reco_d_run1` is 509 km for the smallest test set (5,300 items), it is 168 km for the middle test set (53,000 items). In contrast, *Error* 10 km is considerably more stable, changing from 32.5 to 37.6 % across the two test set sizes.

Third, we also note that despite the random sampling of test cases, some test sets are easier than others—in particular, test set three (53,000) shows the best performance for each run across all evaluated runs. This indicates, that results achieved on particular sub (or super) sets of images cannot be directly compared.

As mentioned above, the run CEA\_run5 shows the influence the use of training data which *includes images from test users* has. *ErrorMedian* now ranges between 1.9 and 2.7 across test cases, while *Error* 10 km ranges from 58 to 63 %. This error magnitude is in line with previous years’ results where training and test users were usually mixed. We stress that although increasing the training corpus also yields small performance gains, the largest gain is achieved by including images of test users in the training data—many users have a particular way of tagging, not only including geographical or in general dictionary terms but they also include long tags (compound phrases meshed together), nicknames or simply tags that only make sense to them. Those tags stand out, and make the location estimation task considerably easier. Thus, we conclude that by partitioning our data set according to training and test users, the task is considerably harder to solve. Exploratory experiments conducted post hoc suggest that after about 2 million sampled training images, the performance gain when adding more training data slowly levels off (assuming that the partitioning of training and test users is intact).

Finally, we turn to the Placeability subtask, which was attempted by one participating group [11]. Similarly to the intuitive example provided earlier, the error is estimated to be large for test images whose location, modeled as a Gaussian, has a high variance. The results of this task, with the linear correlation coefficient varying between 0.06 and 0.37, depending on the evaluated run, indicates that the task itself is feasible. At the same time, the moderate correlation achieved also points to the difficulty and the need for future work: in order to employ such error estimate for any practical means, such moderate correlation is not sufficient.

### 2.2.5 *Placing Task 2014: Horizons*

Currently, MediaEval is preparing to launch the 2014 edition of the Placing Task. Organization of the task was assumed by Bart Thomee, Jaeyoung Choi, Gerald Friedland, and Liangliang Cao. The MediaEval 2014 Placing Task<sup>15</sup> will make use of the Yahoo Flickr Creative Commons 100M dataset.<sup>16</sup> The dataset, also referred to as YLI, was created in collaboration with ICSI Berkeley<sup>17</sup> and the Lawrence Livermore National Laboratory<sup>18</sup> [53]. In this section, we provide a list of highlights that can be expected in the 2014 task.

**Reintroduction of video:** The 2010–2012 editions of the Placing Task focused on geolocation prediction for video, and included both videos and images as training material. The 2013 edition of the task, however, focused on data volume and to this end collected millions of photos for inclusion in the dataset, although at the expense of excluding videos. The exclusion was necessary due to the difficulty in downloading videos in large quantities and the required disk space to store them. In 2014, the task organizers worked to solve download and storage problem. They were prompted to do this by demand among task participants, especially from those that aim to exploit audio and motion features. The 2014 edition will therefore once again feature videos in addition to images, including them both in the training and test sets.

**Evaluation granularity:** In the previous editions of the task, the location predictions were evaluated to be correct within distances of 1, 10, 100, 1,000 and 5,000 km. However, since the accuracy of geotagging has been shown to be much more precise [17] and since higher granularity is a must for various purposes, the 2014 task additionally evaluates the predictions within distances of 10 m and 100 m.

**Russian dolls:** The 2014 test set will be again created according to the “Russian Dolls” approach that was introduced in the 2013 task. This approach divides the test set in multiple sets of varying size each, where each larger set is a superset of all

---

<sup>15</sup> <http://multimediaeval.org/mediaeval2014/placing2014>.

<sup>16</sup> <http://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67>.

<sup>17</sup> <http://www.icsi.berkeley.edu/icsi/>.

<sup>18</sup> <https://www.llnl.gov>.

smaller sets. The Russian Dolls approach enables participants that do not have access to sufficient computational power and/or storage space to pick the largest subset they can handle, while at the same time allowing the evaluation results of all participants to at least be comparable on the smallest subsets they all have in common. This scheme proved successful in the 2013 edition, where one participant used the smallest test set of 5K photos and two participants used second smallest test set of 50K photos, while the remaining four participants used the largest test set of 250K photos.

**Dataset contents:** The dataset of the 2014 edition will be the largest and most complete in comparison with the previous tasks. The collection will contain thousands of videos, millions of photos, device metadata, textual metadata, visual features, and audio features. The dataset will be hosted in the cloud rather than on a single workstation or server to ensure high data availability and fast downloads for the participants.

## 2.3 Future Challenges for Geoprediction of Social Multimedia

The experiences that were gathered over the years of the Placing Task have allowed us to formulate a series of challenges that need to be addressed in order to further improve technology for multimodal geolocation. The ultimate aim of placing algorithms is to generate geolocation information that is truly useful for users within applications. For this reason, we view the challenges of placing as fundamentally determined by the needs of users. In order to ensure that new technologies are directly satisfying human needs, we allow the user perspective to drive our formulation of future challenges for placing.

In this section, we present a picture in which human understanding of the relationship between geolocation and social multimedia gives rise to technical challenges. Our discussion is organized into two parts. The first discusses the different types of relationships that exist between a multimedia item and a location. It uncovers new notions of ‘place’ that need to be taken into account if our systems are to predict geolocations that are maximally meaningful for users. The second discusses the connection between the needs of users and the definition of the Placing Task. Such considerations are necessary in order to ensure that we are addressing a Placing Task that is future proof, in the sense that it will continue to produce technologies that are relevant to user needs, as the use of location-related social multimedia continues to grow.

### 2.3.1 Further Development of Definitions of “Place”

Upon first consideration, defining “place” (i.e., location) seems nearly trivial: it is a set of geocoordinates that encode longitude and latitude. In the following discussion, we will see that there are actually many different notions of place. This variety exists

because concepts of place arise from different sources. One source is geolocation as recorded by the devices that capture multimedia. Another source is human interpretations of geolocation. These sources do not always agree with each other. We discuss each in turn, and then explain why multiple notions of “place” are important for the geolocation prediction for social multimedia, and why it is also critical that we avoid conflating them, but rather that researchers maintain awareness that they are distinct from each other.

### 2.3.1.1 Automatically Captured Geocoordinates

From a technical point of view, the geocoordinates of the camera, or other capture device, used to record a multimedia item provide a straightforward definition of the location of that multimedia item. Information in the form of geocoordinates of the capture device makes an important contribution to the task of placing. It is an easy-to-understand source of location information. More importantly, it allows the problem of geolocation of social multimedia to be studied at large scale, since it is generated without human effort and is associated with a large number of images and videos that are shared by users on-line.

If the performance of geolocation predictions algorithms is to be judged against automatically captured geoinformation, the quality of geocoordinate information is critical so that geolocation prediction algorithms can be reliably assessed. However, limitations exist with respect to the ability of recording devices to capture location information, and media formats to encode location. These limitations can be organized along three dimensions that characterize location description: accuracy, precision, and granularity. Accuracy relates to systematic error between the stated location and the actual location. For automatic systems, accuracy describes how closely the coordinates given by the device match the actual location. The accuracy of the location captured by a recording device can be improved by collecting more readings over longer time periods, or by using more sophisticated signal processing techniques. In the US, the horizontal accuracy of civilian GPS service, Standard Positioning Service (SPS), is often within  $\sim 1$  m.<sup>19</sup> An augmentation system can provide even greater accuracies.<sup>20</sup>

Precision refers to how close individual location readings are to the mean of the reading. The precision of a capture device determines the degree to which a location can be defined (e.g., the number of decimal places in which the geocoordinates can be expressed). It quantifies the random error of geolocation capture. For satellite- and triangulation-based mechanisms, precision is determined by the capability of the system. For placing multimedia within the scope of the entire planet, the precision of geolocation devices is usually sufficient to encode a location such that it distinguishes it from other potential nearby locations. However, for placing with smaller scope,

---

<sup>19</sup> <http://www.gps.gov/systems/gps/performance/accuracy>.

<sup>20</sup> <http://www.gps.gov/systems/augmentations>.

such as being able to accurately distinguish locations inside a building (e.g., statues within a museum), current techniques are not sufficient.

Granularity refers to the geometric correctness-of-the-location description, be it a single point, polygon or polyhedron. The majority of approaches for capturing and encoding location information do so by storing a single geographic point representing the location of the camera. As a result, it is challenging to design placing tasks for social multimedia that move beyond the straightforward assumption that the position of the camera defines the “place” of a photo. Currently, no recording technologies are available that would make possible to create a large-scale social multimedia dataset annotated automatically with geometrically correct descriptions of the locations of objects. Theoretically, it may be desirable to describe an object, and especially a large object like a range of mountains, with bounding rectangle, a planar polygon or even a 3D polyhedron. Whether or not such geolocation descriptions are important for the Placing Task will ultimately depend on whether users find that they provide added value for social multimedia collections.

### 2.3.1.2 Beyond Automatically Captured Geocoordinates

In general, the geocoordinates associated with social multimedia on-line derive from two sources. First, automatically captured geocoordinates, just discussed, and, second, coordinates that are entered by the users at the moment that the images are uploaded. The possibilities for manually associating multimedia with geocoordinates are illustrated by Flickr, which provides users with a map interface. This interface allows them to position their images on a map, which results in an image being assigned geocoordinates, i.e., a geotag. Flickr records the location of photos at 16 levels of accuracy. In the Flickr API documentation, 1 is described as *World*, 3 as *Country*, 6 as *Region*, 11 as *City* and 16 as *Street*.

Location information added by users is dependent on a number of factors. First, it is important to note that the accuracy of manual annotation systems depends on how well the interface maps the provided coordinates to a location on the planet. In the case of manual annotation, the precision is limited only by the interface of the annotation system, be it either a case of adding numerical coordinates directly, or asking a user to select a point on a map. Second, users may be protecting their privacy by choosing to geotag their images at a relatively low level of precision. Third, users may have a limited amount of time and effort to devote to geotagging images, and decide to assign a batch of images to a single location though some of them might not have been taken at that location.

The overall combination of automatically captured geocoordinates and geocoordinates assigned by users results in a wide variation of the characteristics of the geolocation information associated with multimedia. More studies, such as the one on the accuracy of geocoordinate information on Flickr was carried out by [17], are necessary in order to gain a complete understanding of the underlying patterns. Minimally, it is important to be aware that when geotagged images are collected from

Flickr to use to evaluate placing algorithms, the underlying patterns of geotags will impact the measurement of the performance of the system.

Ultimately, the Placing Task will strive to go beyond the geocoordinates that are currently available associated with collections of social multimedia on-line. Specifically, the information should transcend the limitations imposed not only by capture devices, but also by metadata encoding standards. As an example of a common encoding standard, we mention Exchangeable Image File format (Exif), which was initially released in 1995. Exif can encode a wide range of information including technical status values of the capture device (such as aperture, shutter speed for photos) that can be added without human intervention, as well as location information. Currently, however, Exif can encode a single creation location using latitude, longitude and altitude coordinates, as well as a single “destination” location, commonly interpreted as a location pertinent to the content of the image.

Here, we would like to point out several directions that this expansion could take. First, we point out that GPS systems are capable of generating more information than a mere set of geocoordinates. If the error of the capture device is known, it would be helpful to keep information about the error associated with the captured multimedia. When automatically recorded geocoordinates are used as ground truth for evaluating geolocation prediction algorithms, the error limits the resolution with which it is possible to use the ground truth to measure performance.

Second, elevation has an important role to play in defining place. For example, multiple floors in the same building with have the same geolocation, but different elevation. Currently, there is no simply, widely-used manner to add elevation information to social multimedia, either automatically with consumer capture devices, or manually via input interfaces.

Third, it is possible that multiple locations are important for an image. Returning to the example of the mountain range mentioned above, this could be the locations of the individual mountain peaks. Multiple locations are certainly important for video, since the position of the camera may move as the video is recorded. A single set of geocoordinates is not capable of characterizing the capture location of a video, but rather, an entire geopath is necessary.

The catalog of possible descriptions of locations is a long one. Researchers developing placing algorithms need to be able to prioritize the descriptions that they focus on in their research. A great deal of the time, researchers are well served by focusing on using the descriptions of locations that are most readily available at large scale. This point explains the focus in the literature on using the geotags of photos on Flickr as ground truth. Despite the shortcomings of the geoinformation associated with on-line photos mentioned above, there is no other source of ground truth that would allow research datasets of comparable size to be created. However, to the extent to which resources for datasets are available, it is possible to take other priorities into account. Specifically, we focus on descriptions of locations that will best capture the way in which users understand the concept of location as it is related to social images.

### 2.3.1.3 Human Judgements and Interpretations of “Place”

The move towards human views on location starts with the question, “What do we consider to be the *location* of a multimedia item?” Asking this question allows us to uncover *georelevance*, which we define as the connection made by a user between an image or a video and a place. Here, we examine some ways in which people perceive how location is related to multimedia, and reflect on how these correspond to different types of georelevance.

Perhaps the simplest contrast between two types of georelevance is the difference between the location of the camera and the location of the subject material depicted by the multimedia content. This distinction was already mentioned above, in reference to the example of the image of a mountain. We turn to discuss that example again now. The camera that is used to take a picture of the mountain could be positioned kilometers from the mountain itself. It should be noted that the difference in location between the photographer and the subject of the photo is regarded as an inherent fact of photography, and not an exception. It is a common procedure when taking a photo to move away from the subject in order to capture it better. We should be careful in assuming that by default the photograph is at approximately the same location as the subject.

It is easy to ignore the difference between the position of the photographer and the position of the subject. Humans generalize so quickly from different views of objects to the identity of the object themselves. For example, we can recognize the same house taken from different points of view without effort. It is easy to imagine that this interpretation takes place with no conscious awareness of the position of the photographer. This effect can explain why people looking at photos and videos are not specifically aware of the difference in position of the cameraperson and the subject material. That difference is certainly not relevant for interpretation.

More sophisticated notions of relevance require more detailed consideration of the connection between social multimedia and users. Researchers who study social multimedia often study so-called “found” content: content collected from the Internet. It is important to keep in mind that such content is not “found” in a vacuum. Rather, social multimedia comes into existence because it was captured by users. As such, it is natural that a given multimedia item (i.e., an image or a video) will be interpreted differently by different people, and that this different interpretation will also be relevant to place.

In order to understand how the relationship between a person and a multimedia item impacts how that person perceives the connection of that item to location, it is helpful to consider the case of people judging the location of a particular photo. We consider the example of the image in Fig. 2.5, taken in Pisa, Italy.

The relationships that users have with an image determine the evidence and techniques that they have at their disposal for making the decision of the location of that image. Relationships of people to images can be considered to fall into different classes. The classes should not be considered absolute, or necessarily mutually exclusive. However, the existence of these classes shows that different perspectives on location exist among social multimedia users. The following is a list of ten

**Fig. 2.5** Image taken in Pisa, Italy (Flickr: PGBrown1987)



different statements, which could conceivably be made by ten different people (P1–P10), regarding the ways in which they could judge that the location of the image is Pisa.

- P1: I took the picture and when I see it, I remember it.
- P2: I was there when the picture was taken and when I see it, I remember this moment.
- P3: I have been there, and I remember what the place looks like.
- P4: Someone told me about a picture that was taken in Pisa and there is something that I see in this picture that tells me that this must be the picture that the person was talking about.
- P5: I know of another picture that looks just like this one and it was labeled “Pisa”.
- P6: I’ve seen other pictures like this and recognize it (the specific buildings that appear).
- P7: Someone who has been there described this place to me, and on the basis of what we discussed about the place, I recognize it in the picture.
- P8: I’ve been there and recognize characteristics of the place (the type of architecture).
- P9: I am an Internet user, and I can formulate either keyword queries, or use content-based image retrieval to find information that will allow me to identify the picture as Pisa.
- P10: I am a multimedia forensic expert and have established a chain of logic that identifies the place as Pisa.

From these statements, it is clear that a person in class P1, who knows the location of the picture because they took it themselves, will give a different type of answer than a person who is relying on descriptions or other external information. Further, a person who is relying on information will judge the image differently depending on what type of information is available: is it a near duplicate of the same photo, which can be compared side by side, or is it a set of clues concerning details in the pictures that would elude a human judge who was not inclined to detective work.

In the literature, it is conventional to assume a ground truth that is generated by a certain class of users. For example, in formulations of the task of placing that use uploader contributed geocoordinates, the assumption is that users are in class P1. The work of Choi et al. [8], comparing human and machine abilities to generate geopredictions assumed people belonged to class P9. Specifically, in order to collect judgements on human geoprediction performance, crowdsourcing workers were asked to judge images using resources from the Internet, including Google Maps and Streetview.

If different groups of users have different relationships to images, it is not surprising that they also have different notions of the definition of place that is relevant to an image. Another example is helpful in order to communicate this point.

Consider an image of a house. A user could describe this house as “my grandmother’s house”. This definition of place is relative to a particular user. It would be most likely for someone who had taken the picture themselves, or had spent time in the house (approximately corresponding to P1–P4 above). Such a person would judge a geolocation prediction system to have failed if the prediction was off by a few meters (e.g., if the house were assigned the geocoordinates of the neighbor’s house). Another user could describe this house as a “southern bungalow”, referring to a particular architectural style (approximately corresponding to P5–P10 above). This person would judge a geolocation prediction system to have succeeded if it located the neighbors’ house, and quite possibly would be quite satisfied if it correctly placed the image in Florida, USA.

Currently, work in the area of geoprediction conflates these two notions of relevance. It is an open question whether work being done to optimize systems for one type of georelevance directly helps to advance systems optimized for another.

In order to develop detailed and well-supported notions of georelevance, it will ultimately be necessary to carry out careful studies of user interpretations of multimedia location and their location-related multimedia needs. However, we point out already that some findings of such studies may be surprising. For example, it is not unlikely that a large number of users consider an image of a sand dune to be a georelevant depiction of the whole Sahara desert, and that the exact location of the sand dune is relevant only to a minority of users. In this way, geolocation prediction for social multimedia stands in clear contrast with geolocation prediction for military or scientific purposes, in which the exact location may be critical.

Such examples should not automatically be assumed to be exceptions. Social multimedia contains many images which were taken for the purpose of artistic value or emotion impact. Think of a beautiful image of a starry night. An astronomer would be bothered by a small deviation in the accuracy of a geoprediction of the camera position, but a user who is looking for an inspiring or relaxing image would not feel the image to be relevant to any particular geocoordinates. Such considerations would also apply to images of fantastic landscapes, or photographs made as art.

Moving forward, how users interpret place will be strongly impacted by the types of systems that are available for them to interact with geotagged multimedia. As an example of this impact, we consider the case of video. A multimodal location prediction system may be capable of generating a geopath that traces the movement

of the camera during the video. However, unless there is some way to visualize this path for the user, the existence of the path is not useful. A dot tracing the position of a video on the map as the video plays is an interesting solution. However, many questions are left open, such as how to display multiple videos at once.

While allowing more comprehensive geometries to be used to describe a location does not solve this problem, it does give those annotating media better tools for describe location better suited to their media. In addition, explicit annotation for media that have no logical location associated with them would also be valuable. This would help distinguish between media that have no sense of location and those that have not had any location annotated yet.

In summary, users' interpretations of the place that are most relevant for a photo will depend on to which other photos it is being compared, and also on the way in which the user is looking at the photo, or the purpose for which the photo should be used. In the context of annotating multimedia, the existing state of correctness-of-location could possibly be considered sufficient for current needs. However, as technology develops and users begin to demand more from their media collections and systems that handle them, more comprehensive location metadata schema may be required.

Systems will need to be able to serve users who do not present geocoordinates as queries, but rather designate locations by other means. For example, the location of a photo of a person's house may be described by a set of unambiguous coordinates, but also by a postal address, or even a description tied to a specific user, such as "Grandma's House". Multimedia systems need to be able to handle the translation between objective, computer friendly descriptors and subjective, human friendly location descriptors. This would allow multimedia information retrieval systems to take in location queries in forms that are intuitive to humans and be able to search an index of media that have formal location descriptions. The context of the query as well as the personal profile of the user would need to be taken into consideration to deliver effective results.

If the user does not have access to a map that can be referenced for the given lat/lon pair, it is more useful to provide the answer in more practical, abstract or conceptual way such as the postal address or a "tall red building".

### ***2.3.2 Definitions of the Task of Placing Social Multimedia***

An important part of allowing the user perspective to drive our formulation of future challenges for "placing" is understanding the underlying patterns of social multimedia. Social multimedia collections arise due to human behavior, rather than being constructed according to a overall plan or premeditated purpose. For this reason they may be influenced by a range of factors that may not be immediately obvious upon first consideration. Here, we point out that definitions of the Placing Task that do not take these factors into consideration, risk promoting the development of algorithms

that address and artificial problem, whose solutions may not transfer to real-world use scenarios.

A key observation is that different multimedia sharing platforms cater to different communities with different values and, as such, can be considered to have specific cultures. The culture of the platform, for example, includes the extent to which it fosters photography as a hobby, in contrast to other reasons for capturing and sharing images. Also, the technical possibilities of the platform, for example, how easy it is to upload videos, and associate them with tags and geotags, have a large impact on the characteristics of multimedia collections, including patterns of user contributed metadata. Finally, governments impose restrictions around the world on the collection of geotagged multimedia. The definition of what constitutes security risk varies from region to region, and can be expected to have an impact on social multimedia data collections.

The usual reflex of a researcher would be to understand the impact of these factors by making a systematic comparison between multiple datasets and data sources. However, this method of understanding variation between datasets is limited when studying geolocation for social images. As a multimedia sharing platform, Flickr large and so widely used. Other large social multimedia sharing services are able to develop alongside Flickr only to the extent that they serve a different purpose, or supported a different group of users or type of sharing culture. In short, researchers must study social multimedia sharing without having a large number of examples of social multimedia collections of the same type, which would make it possible to formulate a set of properties characteristic of a specific social multimedia collection. The study of social multimedia is the study of specific collections, and not the study of an ideal.

For this reason, it is important when creating a dataset that is to be used to develop geolocation prediction algorithms, that the underlying characteristics of the social multimedia collection from which it is drawn are preserved as much as possible. In other words, dataset development should leave “found data” as much as possible in the context in which it was found. The more that we change the data from its “in the wild” state, the less certain we can be that we are developing algorithms, or applications, that are relevant to multimedia that users actually produce and are appropriate for the patterns with which they produce it.

Unfortunately, it is easy to get frustrated with “found data” because its properties are a priori unknown and out of control of the researcher. The natural temptation is to seek to somehow “improve” it. This temptation arises from our drive, as researchers, to have the best possible dataset on which to carry out our research. It is critical, that we define “best possible” dataset by considering the ultimate goal of our research, and not focusing on what characteristics that our algorithms require in a dataset in order to perform well. However, before we begin to devote effort to address a particular “drawback” or “shortcoming” of our dataset, it is important to give careful consideration to whether or not that shortcoming is a reflex of the underlying problem to be solved.

An example serves to illustrate issues that can arise when we are overhasty to identify an aspect of a dataset as a weakness. If we are tackling the goal of creating

a better search and browsing functionalities for users of social multimedia websites, we may feel that one “shortcoming” of the dataset is the fact that people take pictures inside buildings, which possibly contain very little information distinctive for location. If we “improve” our dataset by discarding all images taken inside buildings, we actually have changed that underlying problem. We will no longer be working toward algorithms that have a chance of improving search and browsing functionalities overall, but rather only for a portion of all user-uploaded images. Changing the dataset changes important aspects such as the prior probability of certain locations, the composition of the negative class, and the relative proximity of images to each other in feature spaces, such as the visual feature space. In practice, difficult challenges may be tackled by isolating subchallenges and addressing them individually. However, the overall implications of simplifying the dataset must be carefully considered.

We can formulate best practice with the following statement: Researchers should decide on the ultimate goal of the research independently of the collection of the dataset, and where every possible datasets should be collected “in the wild” from the communities that are ultimately meant to benefit from the algorithms that researchers develop. Decisions to change the distribution of a dataset should not be viewed as “improvements,” but rather should clearly be given the status of informed modifications of the original data, carried out with the purpose of accomplishing a subgoal. Note that because the multimodal geolocation requires such large amounts of data, the decisions to “change” datasets from their naturally occurring state does not take place after the data has been collected, but rather during the collection process. It is important to reflect thoroughly on whether the methodology used for crawling or sampling pushes the dataset away from the “use case” corresponding to the ultimate goal of the research.

## 2.4 Conclusion and Outlook

This article has presented a retrospective on the Placing Task offered by the Media-Eval Multimedia Benchmark. We have traced the first four years of the benchmark 2010–2013, and provided an outlook on 2014. In this final section, we summarize essential points concerning the ground covered by the task thus far, and anticipate the new challenges that it will tackle in the immediate future.

### 2.4.1 *Where Placing has Been*

The Placing Task has successfully provided large-scale datasets to the multimedia community that have supported cross-site comparison of multimodal location prediction algorithms.

Several important principles have emerged that transcend specific algorithms. One is that two-stage approaches are highly effective. The first stage creates a set of candidate hypotheses and the second stage seeks to refine these hypotheses. The first two-stage approach was introduced by Olivier Van Laere in 2010 [58] and has also been exploited by Michele Trevisiol and colleagues [56]. Another principle is that the test data can be jointly exploited to address issues of data sparsity. Both Jaeyoung Choi and colleagues in 2012 and Jiewei Cao in 2013 exploited this principle within two different algorithms. Finally, the importance of user modeling has been made clear, both by approaches that build models of past tagging behavior and approaches that default to users' home locations. Here, notable contributions include [7] in 2010, [60] in 2011, and [48]. In 2013, the task was designed to encourage participants to emphasize aspects other than the same-uploader signal.

The MediaEval Placing Task has successfully guided researchers to explore new areas. In 2010, only a single group (TU Berlin [24]) made use of visual features alongside image metadata. By 2013, over half of the submitted runs used a combination of textual and visual features, or visual features alone. Likewise, the task has successfully guided researchers away from less promising areas. Already in 2010, [7] and the follow-up work [14] reveals that the use of knowledge resources such as gazetteers that contain lists of geographically related words, are not the silver bullet for addressing the task of geolocation social multimedia. Instead, if enough data is available, data-driven approaches will outperform approaches that use knowledge resources. As a result, MediaEval participants used gazetteers judiciously, and did not automatically assume that knowledge resources were necessary to implement effective text-based geolocation prediction approaches.

Other results achieved by MediaEval involve the research community as a whole, rather than individual algorithms. The fact that researchers worked together within the framework of a benchmark rather than as individuals has strengthened the research in this area. The Placing Task 2010–2013 demonstrates that researchers can successfully improve their approaches via discussion with other teams at the yearly workshop, and results on extended algorithms have gone on to be published in mainstream venues.

The benchmarking framework is also important in driving forward the state of the art. Specifically, the Placing Task experience has shown that encouraging state-of-the-art baseline works: When provided with specific papers and code, participants did better at starting where others had left off, rather than re-inventing the wheel. Lowering the threshold works: Sites could participate that would not otherwise have been able to take part, had they not had access to data, features, and also had the “Russian Doll” approach allowed them to work on a smaller dataset. Students benefit from the support of the community as they carry out their thesis research.

Finally the Placing Task has brought benefits to the research community by opening the door to exploration of new topics. It is a laudable goal to make social multimedia more easily findable and browsable by annotating it with location information. However, it is also critical to remember that geoinformation does not always make a positive contribution. Many Internet users avoid geotagging their videos and photos, or turn the GPS functionality of their phones off, since they do not want the location of their media items to be known. An important development was that ICSI was able

to use the MediaEval 2010 dataset in order to move forward with its work in the area of privacy [14].

### ***2.4.2 Where Placing Is Going***

As discussed in Sect. 2.3, researchers developing automatic geoprediction algorithms for social multimedia stand before a large number of challenges. We have argued that the most worthwhile future challenges are those driven by user needs or user behavior.

We have discussed a variety of notions of georelevance that arise from how people interpret images. These new notions of georelevance will also require new evaluation metrics. Currently, the Placing Task evaluates the quality of a prediction by calculating the Haversine distance between the predicted geocoordinates and the ground truth. Distance does not necessarily correspond to human perceptions of relevance, however. We have already mentioned the importance of the relationship of the person to the multimedia item in determining their perception of georelevance. However, the problem cannot be solved by merely enlarging the evaluation radius, but rather researchers must dig deeper.

One effect that they will uncover is that human interpretations of georelevance can be assumed to be spatially discontinuous. Users may consider an image to be accurately geolocated if its location is predicted to be anywhere along the shore of a lake. However, if the location is predicted to be in the middle of the lake, users would feel that the geolocation is inaccurate. Such judgements are clearly related to the semantic content of the example.

In the future the Placing Task will need to decide the extent to which it is addressing a challenge of image understanding. When humans interpret images, it is possible for them to find image content to be relevant to a geolocation without the image content being physically located at the geolocation. Clearly, if a geolocation prediction system identifies a picture of a cowboy in New York City as “Texas” would be less disturbing for a user than geotagging the Statue of Liberty “Texas”. User studies are necessary to understand the extent to which users expect and are able to use systems that can predict geoconnections going beyond physical location.

In any case, future work will necessarily involve a better understanding of the properties of social multimedia data that allow geocoordinate prediction systems to work. For example, in the case of visual features, it remains unclear if systems should be optimized to identify near duplications, to match scenes, to match specific landmarks, or to match types of objects. This choice will determine the types of visual features that are best suited to the task.

Overall, geoprediction stands to benefit from a better understanding of when particular approaches are most likely to work well and when they should be avoided. For example, currently, many systems apply the same approach to predicting the geolocation of multimedia items in regions that are represented with a wealth of multimedia content to multimedia items where data is sparse. Instead, a system

could first predict whether it is confronting a “easy” or “hard” prediction, and then react accordingly. The initial groundwork has been laid for such approaches by the Placing Task 2013 Placeability subtask. Further, the use of audio for predicting the geolocation of social video has delivered some initially promising results. Here, it is still necessary to understand the types of situations in which audio can be used to the best advantage.

In all cases, algorithms must anticipate the challenges they face as the amount of available social multimedia grows larger and perhaps even changes in composition. In this chapter, we have focused heavily on Flickr. This focus is primarily dictated by the availability of the data for research purposes. However, many other sources of online multimedia exist, including: YouTube,<sup>21</sup> Facebook,<sup>22</sup> Vimeo,<sup>23</sup> blip.tv,<sup>24</sup> Instagram,<sup>25</sup> and Vine.<sup>26</sup> New patterns in multimedia collections can be expected as new capture devices are introduced (e.g., sports cameras, Google Glass) and users change their capture and sharing habits.

In the face of increasing amounts of multimedia data, it is important to remember that ultimately we are best served by not only reflecting on where we can obtain more data, but also, how we can be sure that we are using the right data. Ultimately, the Placing Task must be steered by the geoprediction needs of users for multimedia-associated geoinformation. We should be alert to cases in which users needs can be addressed by a smaller number of specific multimedia items. Such cases could be expected to arise, for example, in geolocation systems for constrained location, i.e., geolocating images taken within a museum.

Moving forward, it will be important to pursue new avenues of research that are opened by the MediaEval Placing Task. As already mentioned above, developing algorithms that are capable of automatically prediction the geolocation of social multimedia has important implications for user privacy. In view of these implications, it is important to study not only the geolocation problem, but also the inverse challenge. Addressing this challenge involves determining if it possible to maintain those properties that make multimedia items understandable and interesting to users, while at the same time hiding the information necessary to produce geopredictions. We refer to this challenge as “Geo-Cloaking”. A geocloaking task could involve merging the Placing Task with another task offered at MediaEval, namely Visual Privacy [2]. The Visual Privacy task seeks to develop methods of obscuring video that are acceptable to human viewers, while at the same time also obscuring person information. In order to test the effectiveness of such systems, state-of-the-art geolocation prediction algorithms are necessary.

Another interesting avenue is to turn the Placing Task on its head. Instead of inferring information about an multimedia item in the form of a location, it is possible

---

<sup>21</sup> <https://www.youtube.com>.

<sup>22</sup> <https://www.facebook.com>.

<sup>23</sup> <https://vimeo.com>.

<sup>24</sup> <http://blip.tv>.

<sup>25</sup> <http://instagram.com>.

<sup>26</sup> <https://vine.co>.

to conceptualize the task as inferring information about a location by gathering information about multimedia items. Ghent University has taken the lead in work dedicated to learning geographically relevant information by using georeferenced social media [30]. This perspective is an interesting one for future pursuit.

We close this chapter with a few words on the larger implications of the MediaEval Placing Task. We have seen in this article that the Placing Task represents a multi-year, international cooperative effort to solve the overall problem of predicting the geocoordinates of a social multimedia item, wherever in the world that item might be located. It distinguishes itself from other initiatives in the area of geocoordinate prediction by its focus on large collections of social multimedia—it seeks to solve the problem using “found” datasets gathered from online multimedia sharing platforms and transferred as directly as possible into a task. The most advanced resource imaginable to tackle this problem would be a multimedia collection containing every item available in the world. If individual research sites were to develop such a collection, few would be able to afford to carry out Placing research. The Placing Task can, for this reason, be considered to share similarities with future-oriented initiatives that are formulated on a grand scale, such as the International Space Station.<sup>27</sup> Such projects require an international coalition which comes together to build and maintain a resource that allow scientific research to be carried out that would not be possible any other way. The Placing Task has made a strong start, and its organizers look forward to further growth in the future. As they push forward the state of the art in geolocation of social multimedia, the sky is the limit.

## References

1. J. Almeida, N. Leite, R. Torres, Comparison of video sequences with histograms of motion patterns, in *18th IEEE International Conference on Image Processing (ICIP)*, September 2011, pp. 3673–3676
2. A. Badii, M. Einig, T. Piatrik, Overview of the MediaEval 2013 Visual Privacy Task, in Larson et al. [31]
3. J. Cao, Photo set refinement and tag segmentation in georeferencing Flickr photos, in Larson et al. [31]
4. J. Choi, V. Ekambaram, G. Friedland, K. Ramchandran, The 2012 ICSI/Berkeley video location estimation system, in Larson et al. [35]
5. J. Choi, G. Friedland, Data-driven vs. semantic-technology-driven tag-based video location estimation, in *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing, ICSC '11*. IEEE Computer Society, Washington, DC, pp. 243–246 (2011)
6. J. Choi, G. Friedland, V. Ekambaram, K. Ramchandran, Multimodal location estimation of consumer media: dealing with sparse training data, in *Proceedings of the 2012 IEEE International Conference on Multimedia and Expo, ICME '12*. IEEE Computer Society, Washington, DC, pp. 43–48 (2012)
7. J. Choi, A. Janin, G. Friedland, The 2010 ICSI video location estimation system, in Larson et al. [33]
8. J. Choi, H. Lei, V. Ekambaram, P. Kelm, L. Gottlieb, T. Sikora, K. Ramchandran, G. Friedland, Human versus machine: establishing a human baseline for multimodal location estimation, in

---

<sup>27</sup> [http://www.nasa.gov/mission\\_pages/station](http://www.nasa.gov/mission_pages/station).

- Proceedings of the 21st ACM International Conference on Multimedia, MM '13*, ACM, New York, pp. 867–876 (2013)
9. J. Choi, H. Lei, G. Friedland, The 2011 ICSI video location estimation system, in Larson et al. [32]
  10. D.J. Crandall, L. Backstrom, D. Huttenlocher, J. Kleinberg, Mapping the world's photos, in *Proceedings of the 18th International Conference on World Wide Web, WWW '09*, ACM, 2009, pp. 761–770
  11. J. Davies, J. Hare, S. Samangoei, J. Preston, N. Jain, D. Dupplaw, P. Lewis, Identifying the geographic location of an image with a multimodal probability density function, in Larson et al. [31]
  12. D. Ferrès, H. Rodríguez, TALP at MediaEval 2010 Placing Task: geographical focus detection of Flickr textual annotations, in Larson et al. [33]
  13. D. Ferrès, H. Rodríguez, TALP at MediaEval 2011 Placing Task: georeferencing Flickr videos with geographical knowledge and information retrieval, in Larson et al. [32]
  14. G. Friedland, J. Choi, Semantic computing and privacy: a case study using inferred geo-location. *Int. J. Semant. Comput.* **5**(1), 79–93 (2011)
  15. G. Friedland, J. Choi, A. Janin, VIDEO2GPS: a demo of multimodal location estimation on Flickr videos, in *Proceedings of the 19th ACM International Conference on Multimedia, MM '11*, ACM, New York, pp. 833–834 (2011)
  16. A. Gallagher, D. Joshi, J. Yu, J. Luo, Geo-location inference from image content and user tags, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2009, CVPR Workshops 2009*, June 2009, pp. 55–62
  17. C. Hauff, A study on the accuracy of Flickr's geotag data, in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13*, ACM, New York, pp. 1037–1040 (2013)
  18. C. Hauff, G.-J. Houben, WISTUD at MediaEval 2011: placing task, in Larson et al. [32]
  19. C. Hauff, G.-J. Houben, Geo-location estimation of Flickr images: social web based enrichment, in *Proceedings of the 34th European Conference on Advances in Information Retrieval, ECIR'12*. Springer, Berlin, pp. 85–96 (2012)
  20. C. Hauff, G.-J. Houben, Placing images on the world map: a microblog-based enrichment approach, in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12*, ACM, New York, pp. 691–700 (2012)
  21. C. Hauff, B. Thomee, M. Trevisiol, Working notes for the placing task at MediaEval 2013, in Larson et al. [31]
  22. J. Hays, A.A. Efros, Im2gps: estimating geographic information from a single image, in *CVPR*. IEEE Computer Society (2008)
  23. J.M. Perea-Ortega, M.Á. García-Cumbreras, L. Alfonso Ureña-López, M. García-Vega, SINAI at Placing Task of MediaEval 2010, in Larson et al. [33]
  24. P. Kelm, S. Schmiedeke, T. Sikora, VIDEO2GPS: geotagging using collaborative systems, textual and visual features: MediaEval 2010 Placing Task, in Larson et al. [33]
  25. P. Kelm, S. Schmiedeke, T. Sikora, A hierarchical, multi-modal approach for placing videos on the map using millions of Flickr photographs, in *ACM Multimedia 2011 (Workshop on Social and Behavioral Networked Media Access—SBNMA)*, ACM, November 2011
  26. P. Kelm, S. Schmiedeke, T. Sikora, Multi-modal, multi-resource methods for placing Flickr videos on the map, in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ICMR '11*, ACM, New York, pp. 52:1–52:8 (2011)
  27. P. Kelm, S. Schmiedeke, T. Sikora, How spatial segmentation improves the multimodal geotagging, in Larson et al. [35]
  28. G. Kordopatis-Zilos, S. Papadopoulos, E. Spyromitros-Xioufis, A.L. Symeonidis, Y. Kompatsiaris, CERTH at MediaEval Placing Task 2013, in Larson et al. [31]
  29. F. Krippner, G. Meier, J. Hartmann, R. Knauf, Placing media items using the XTRetrieval framework, in Larson et al. [32]
  30. O.V. Laere, S. Schockaert, V. Tanasescu, B. Dhoedt, C. Jones, Georeferencing Wikipedia documents using data from social media. *ACM Trans. Inf. Syst.* **32**(3), (2014)

31. M. Larson, X. Anguera, T. Reuter, G.J.F. Jones, B. Ionescu, M. Schedl, T. Piatrik, C. Hauff, M. Soleymani (eds.), in *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2013, CEUR-WS.org, online <http://ceur-ws.org/Vol-1043> (2013)
32. M. Larson, A. Rae, C.-H. Demarty, C. Kofler, F. Metze, R. Troncy, V. Mezaris, G.J.F. Jones (eds.), in *Working Notes Proceedings of the MediaEval 2011 Workshop*, Pisa, Italy, September 2011, CEUR-WS.org, online <http://ceur-ws.org/Vol-807> (2011)
33. M. Larson, M. Soleymani, P. Serdyukov, V. Murdock, G.J.F. Jones (eds.), in *Working Notes Proceedings of the MediaEval 2010 Workshop*, Pisa, Italy, October 2010, online <http://multimediaeval.org/mediaeval2010/2010worknotes> (2010)
34. M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, G.J.F. Jones, Automatic tagging and geotagging in video collections and communities, in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ICMR '11*, ACM, New York, pp. 51:1–51:8 (2011)
35. M. Larson, S. Schmiedeke, P. Kelm, A. Rae, V. Mezaris, T. Piatrik, M. Soleymani, F. Metze, G.J.F. Jones (eds.), in *Working Notes Proceedings of the MediaEval 2012 Workshop*, Pisa, Italy, October 2012, CEUR-WS.org, online <http://ceur-ws.org/Vol-927> (2012)
36. M. Larson, M. Soleymani, M. Eskevich, P. Serdyukov, R. Ordelman, G. Jones, The Community and the Crowd: Multimedia Benchmark Dataset Development. *MultiMedia*, IEEE. **19**(3), 15–23 (2012)
37. H. Lei, J. Choi, G. Friedland, Multimodal city-verification on Flickr videos using acoustic and textual features, in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 2273–2276
38. L. Li, D. Pedronette, J. Almeida, O. Penatti, R. Calumby, R. Torres, A rank aggregation framework for video multimodal geocoding, pp. 1–37 (2013)
39. L.T. Li, J. Almeida, R.D.S. Torres, RECOD working notes for placing task MediaEval 2011, in Larson et al. [32]
40. L.T. Li, J. Almeida, D.C.G. Pedronette, O. Penatti, R.D.S. Torres, A multimodal approach for video geocoding, in Larson et al. [35]
41. L.T. Li, J. Almeida, O. Penatti, R. Calumby, D.C.G. Pedronette, M.A. Gonçalves, R.D.S. Torres, Multimodal image geocoding: the 2013 RECOD's approach, in Larson et al. [31]
42. X. Li, C. Hauff, M.A. Larson, A. Hanjalic, Preliminary exploration of the use of geographical information for content-based geo-tagging of social video, in Larson et al. [35]
43. X. Li, M. Riegler, M. Larson, A. Hanjalic, Exploration of feature combination in geo-visual ranking for visual content-based location prediction, in Larson et al. [31]
44. N. O'Hare, V. Murdock, Modeling locations with social media. *Inf. Retr.* **16**(1), 30–62 (2013)
45. J. Oomen, P. Over, W. Kraaij, A. Smeaton, Symbiosis between the TrecVid benchmark and video libraries at the Netherlands Institute for Sound and Vision. *Int. J. Digit. Libr.* **13**(2), 91–104 (2013)
46. O.A.B. Penatti, L.T. Li, J. Almeida, R.D.S. Torres, A visual approach for video geocoding using bag-of-scenes, in *Proceedings of the 2Nd ACM International Conference on Multimedia Retrieval, ICMR '12*, ACM, New York, pp. 53:1–53:8 (2012)
47. A. Popescu, CEA List's participation at MediaEval 2013 Placing Task, in Larson et al. [31]
48. A. Popescu, N. Ballas, CEA List's participation at MediaEval 2012 Placing Task, in Larson et al. [35]
49. A. Rae, P. Kelm, Working notes for the Placing Task at MediaEval 2012, in Larson et al. [35]
50. A. Rae, V. Murdock, P. Serdyukov, P. Kelm, Working notes for the Placing Task at MediaEval 2011, in Larson et al. [32]
51. S. Schmiedeke, C. Kofler, I. Ferrané, Overview of the MediaEval 2012 Tagging Task, Working Notes Proceedings of the MediaEval 2012 Workshop, Santa Croce in Fossabanda, Pisa, Italy, October 4–5, CEUR-WS.org, ISSN 1613–0073 (2012)
52. P. Serdyukov, V. Murdock, R. van Zwol, Placing Flickr photos on a map, in *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09*, ACM, New York, pp. 484–491 (2009)

53. D.A. Shamma, One hundred million creative commons Flickr images for research. <http://yahoolabs.tumblr.com/post/89783581601/one-hundred-million-creative-commons-flickr-images-for>, month = June, note = Accessed: 30 June 2014 (2014)
54. A.F. Smeaton, P. Over, W. Kraaij, Evaluation campaigns and TrecVid, in *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, MIR '06*, ACM, New York, pp. 321–330 (2006)
55. S. Subramanian, V. Vidyasagan, K. Chandramouli, VIT@MediaEval 2013 Placing Task: location specific tag weighting for language model based placing of images, in Larson et al. [31]
56. M. Trevisiol, J. Delhumeau, H. Jégou, G. Gravier, How INRIA/IRISA identifies geographic location of a video, in Larson et al. [35]
57. M. Trevisiol, H. Jégou, J. Delhumeau, G. Gravier, Retrieving geo-location of videos with a divide & conquer hierarchical multimodal approach, in *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval, ICMR '13*, ACM, New York, pp. 1–8 (2013)
58. O. Van Laere, S. Schockaert, B. Dhoedt, Ghent University at the 2010 Placing Task, in Larson et al. [33]
59. O. Van Laere, S. Schockaert, B. Dhoedt, Finding locations of Flickr resources using language models and similarity search, in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ICMR '11*, ACM, New York, pp. 48:1–48:8 (2011)
60. O. Van Laere, S. Schockaert, B. Dhoedt, Ghent University at the 2011 Placing Task, in Larson et al. [32]
61. O. Van Laere, S. Schockaert, B. Dhoedt, Georeferencing Flickr photos using language models at different levels of granularity: an evidence based approach. *J. Web Semant.* **16**, 17–31 (2012)
62. O. Van Laere, S. Schockaert, B. Dhoedt, Georeferencing Flickr resources based on textual meta-data. *Inf. Sci.* **238**, 52–74 (2013)
63. O. Van Laere, S. Schockaert, J. Quinn, F. Langbein, B. Dhoedt, Ghent and CARDIFF University at the 2012 Placing Task, in Larson et al. [35]



<http://www.springer.com/978-3-319-09860-9>

Multimodal Location Estimation of Videos and Images

Choi, J.; Friedland, G. (Eds.)

2015, XII, 191 p. 80 illus. in color., Hardcover

ISBN: 978-3-319-09860-9