Geotagging Text Content With Language Models and Feature Mining

Article *in* Proceedings of the IEEE · August 2017

DOI: 10.1109/JPROC.2017.2688799

Giorgos Kordopatis-Zilos
The Centre for Research and Technology, Hellas

Symeon Papadopoulos
The Centre for Research and Technology, Hellas

Ioannis (Yiannis) Kompatsiaris
The Centre for Research and Technology, Hellas

Geotagging Text Content with Language Models and Feature Mining

Giorgos Kordopatis-Zilos, *Student Member, IEEE,* Symeon Papadopoulos, *Member, IEEE,* Ioannis (Yiannis) Kompatsiaris, *Senior Member, IEEE*

***Abstract*—The large-scale availability of user-generated content in social media platforms has recently opened up new possibilities for studying and understanding the geospatial aspects of real-world phenomena and events. Yet, the large majority of user-generated content lacks proper geographic information (in the form of latitude and longitude coordinates). As a result, the problem of multimedia geotagging, i.e. extracting location information from user-generated text items when this is not explicitly available, has attracted increasing research interest. Here, we present a highly accurate geotagging approach for estimating the locations alluded by text annotations based on refined language models that are learned from massive corpora of social media annotations. We further explore the impact of different feature selection and weighting techniques on the performance of the approach. In terms of evaluation, we employ a large benchmark collection from the MediaEval Placing Task over several years. We demonstrate the consistently superior geotagging accuracy and low median distance error of the proposed approach using various datasets and comparing it against a number of state-of-the-art systems.**

***Index Terms*—geotagging, geolocation, language model, feature selection, location estimation**

I. Introduction

The ubiquitous availability and use of media capturing devices (smartphones, cameras) and the increasing penetration of online social networking and media sharing services have led to massive increase in the amount of user-generated content and discussions related to unfolding news stories and real-world events. A key element of user-generated content is text, which constitutes either the only component of a social media post, e.g. a tweet, or an annotation that accompanies a multimedia item (e.g. Flickr image, YouTube video). Text annotations are also an important part of online user profiles (e.g. the description field of a Twitter account). In many cases, user-generated text annotations are indicative of the location they originate from or they refer to, either because they explicitly mention particular geographic entities or because the text contains cues that implicitly refer to particular locations.

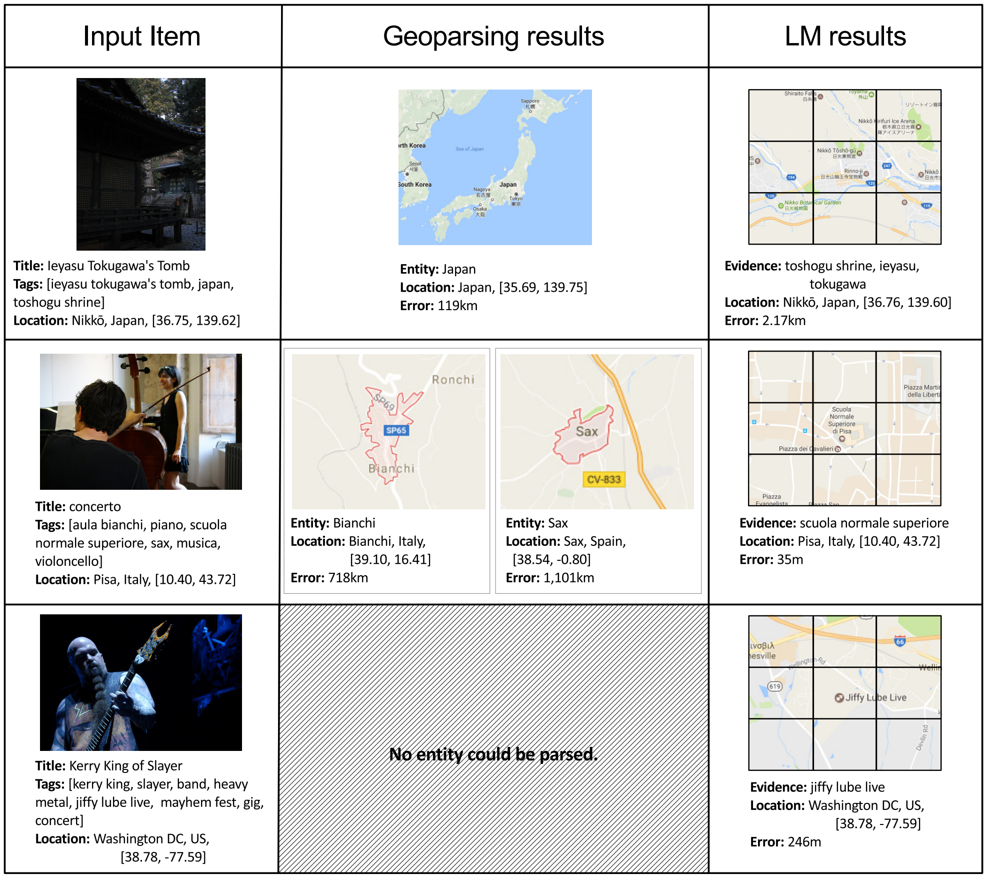
Yet, the majority of user-generated text is not accompanied by proper geographic information either due to the way they are generated (e.g., the mobile app used to upload a tagged image is set to not use the GPS coordinates of the device) or due to social media platform policies (for instance, Facebook and Twitter remove all Exif metadata, including geolocation, from any uploaded image). Being able to perform geotagging is often very valuable for social media monitoring applications. For instance, a journalist may be able to identify and cross-check the location of a breaking news item by corroborating multiple social media posts that have been automatically geotagged. In another example, a business analyst may be able to determine the popularity of a particular brand over the globe by aggregating the locations that have been automatically extracted from social media accounts that mention the brand in their posts. To this end, the problem of geotagging has attracted increasing research interest and a variety of geotagging methods have been proposed to tackle it [36], [56].

The most widely used approach for geotagging is geoparsing, i.e. the detection of references to known locations with the help of a gazetteer [1]. Yet, it is often extremely challenging to reliably infer the location alluded by an arbitrary piece of text content using such approaches due to the inherent complexity of the problem and the nature of social media content. In particular, there is a number of limitations and challenges faced by geoparsing methods (Figure 1):

- In addition to "fixed" and well-known location names (e.g. cities, neighbourhoods, landmarks), there is a huge number of geographic names that are often dynamic (e.g. shop names, emerging new hip areas). As a result, maintaining a comprehensive and up-to-date list of geographic names is a very challenging task.
- More often than not, it is often possible to infer the location of a text description without any explicit reference to geographic entities. Such cases appear for instance when there are references to local food or when there is use of local slang.
- There is often considerable ambiguity induced by geographic names, since the same name may refer to different locations in the world. For instance, the name `Athens` refers to more than 20 locations in the US in addition to the capital of Greece. Thus, the presence of a single geographic name without any additional context may be misleading in terms of the implied location.

To this end, another popular alternative to geotagging is the use of Language Model-based (LM) approaches [43]. Those attempt to learn probabilistic text models, which given a piece of text, provide an estimate of the likelihood that the text refers to a particular location. LM approaches hold the promise of alleviating a number of the above challenges faced by geoparsing methods, since they do not operate on the basis of an explicit toponym dictionary, but are instead trained on large

G. Kordopatis-Zilos, S. Papadopoulos and Y. Kompatsiaris are with the Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH), 6th km Charilaou - Thermi, 57001, Thessaloniki, Greece. e-mails: {georgekordopatis, papadop, ikom}@iti.gr



Input Item	Geoparsing results	LM results
Title: Ieyasu Tokugawa's Tomb **Tags:** [ieyasu tokugawa's tomb, japan, toshogu shrine] **Location:** Nikkō, Japan, [36.75, 139.62]	**Entity:** Japan **Location:** Japan, [35.69, 139.75] **Error:** 119km	**Evidence:** toshogu shrine, ieyasu, tokugawa **Location:** Nikkō, Japan, [36.76, 139.60] **Error:** 2.17km
Title: concerto **Tags:** [aula bianchi, piano, scuola normale superiore, sax, musica, violoncello] **Location:** Pisa, Italy, [10.40, 43.72]	**Entity:** Bianchi **Location:** Bianchi, Italy, [39.10, 16.41] **Error:** 718km **Entity:** Sax **Location:** Sax, Spain, [38.54, -0.80] **Error:** 1,101km	**Evidence:** scuola normale superiore **Location:** Pisa, Italy, [10.40, 43.72] **Error:** 35m
Title: Kerry King of Slayer **Tags:** [kerry king, slayer, band, heavy metal, jiffy lube live, mayhem fest, gig, concert] **Location:** Washington DC, US, [38.78, -77.59]	**No entity could be parsed.**	**Evidence:** jiffy lube live **Location:** Washington DC, US, [38.78, -77.59] **Error:** 246m

Fig. 1. Visual examples of the estimated locations based on a state-of-the-art geoparsing method [55] (second column) and our proposed Language Model-based approach (third column). In the first case, the geoparsing approach only managed to detect a very coarse entity (Japan) close to the true location referred in the text of the input image. In the second case, the geoparsing method was confused due to the ambiguity of the mentioned location. In the third case, the geoparsing method could not detect any location. Instead, our proposed LM-based approach correctly estimated the true location of all three input images.

corpora of geotagged text, which are nowadays relatively easy to collect and are rich in terms of location-specific language. In addition, being probabilistic in nature, LM approaches take text context into account and do not rely on a specific term-entity to produce a reliable estimate. However, many LM-based approaches suffer from two weaknesses: a) they are tuned to produce good estimates at a given *geographic granularity*, e.g. region, city or at best neighbourhood, and often suffer from lack of precision or robustness (i.e. in case of an error, their estimate may be very far from the true location); b) they are sensitive to the training set used to generate the LM and often end up being dataset-specific, i.e. they suffer from overfitting.

In this paper, we address the limitations of previous LM-based approaches to deliver a highly accurate and robust geotagging approach. To this end, we propose two important extensions over previous high-performing LM-based systems:

- Unlike previous LM-based approaches that rely on a *single grid of cells*, such as the geotagging system by Popescu [40] (which achieved the best results in the 2013 edition of the MediaEval Placing Task), we propose the use of *multiple grids* to capture language geographicity at different granularities, in a way that ensures both precise geotagging (i.e. providing estimates that are as close as possible to the true locations) and resilience in cases where this is not possible (i.e. provide the best possible geotagging estimate, avoiding very large errors). As a result, the precision at low granularities improved almost 10 times. More details on the employed methods are presented in Sections III-B and III-E.
- Extending previous approaches that use feature selection during the LM construction step, such as our previous approach [23], we propose a more versatile, scalable and powerful feature selection and weighting scheme, which leads to considerable improvement in terms of geotagging accuracy and to increased resilience with respect to the training dataset. In particular, the impact of feature selection is the reduction of median distance error up to ≈88%. The proposed feature selection and weighting schemes were applied on a training set of ≈40M geotagged text annotations on a commodity server. More details on the feature selection and weighting methods are presented in Sections III-C and III-D.

Moreover, we present a comprehensive experimental study (in Section IV) using the YFCC100M large-scale dataset [48] and four editions of the MediaEval Placing Task (2013-2016), in all of which the proposed method achieved the best or very close to best performance. In the latest edition (2016), our method achieved top performance, with $P@1km$=24.85% and median error equal to 28km, which was further improved to 27.4% and 16km respectively, when training the method with the full YFCC dataset. This is the highest reported geotagging accuracy in the history of the benchmark when using a completely independent training dataset and no external resources (e.g. gazetteers). In addition, we evaluate the contribution of a number of state-of-the-art techniques [23], [50], [51], as well as of increasing the size of the training set, to the overall performance of the geotagging process. To further drive research in the area, we publish the implementation of the proposed approach as an open-source project[1].

II. Related Work

Geotagging social media content is a challenging task, which has attracted increasing research interest in recent years. Detailed surveys of the field were presented in [36] and [56], discussing a variety of geotagging approaches. Text-based approaches are classified into two broad categories: *geoparsing* and *Language Model*-based (LM). Geotagging approaches based on the visual content of images, such as the ones by Hayes et al. [16], [17], Lin et al. [35], Weyand et al. [54] and Li et al. [32], offer another interesting alternative solution to the problem, which is, however, beyond the scope of this paper. Similarly, multimodal approaches that combine both text and visual content to produce location estimates, such as the ones by Crandall et al. [9], Kelm et al. [20], Trevisiol et al. [49] and Cao et al. [3] are not further considered here. Yet, it is noteworthy that text-based geotagging approaches are currently much more accurate and reliable compared to visual-based ones, while combined approaches have at the moment only marginal gains compared to text-based approaches at considerable added complexity. For instance, in MediaEval 2015, the best text-based submission achieved a score of $P@1km$ = 27.3%, the best visual only a score of $P@1km$ = 5.2%, while the best combined approach a score of $P@1km$ = 27.54% [26]. In contrast, approaches that exploit information about the author/creator of a social media post can achieve massive gains in performance. For instance, Popescu et al. [41] achieved the best result in MediaEval 2014 ($P@1km$ = 44.13% and m.error=1.9km), by taking into account the recent locations (past 24 hours) of the Flickr users that uploaded the test images. However, such methods are only applicable in limited scenarios, and thus fall outside the scope of our research.

A. Geoparsing

Gazetteers are essentially large dictionaries or directories that contain comprehensive lists of geographic entities. These are described by various features, such as location, toponym and alternate names (when available). Gazetteers typically contain high quality and precise information. However, many of them have limited world coverage, which makes them insufficient as a basis for a global geotagging solution. The most well-known gazetteers are Geonames[2], OpenStreetMap[3] and Yahoo! GeoPlanet[4] and DBpedia [28] (which is not limited to geographical entities).

Several geotagging approaches are based on gazetteers. One of the earlier works in the field was presented by Amitay et al. [1]. This combined different gazetteers to determine the locations of mentioned places in web content. Keßler et al. [21] combined existing standards to realize a gazetteer infrastructure allowing for bottom-up contribution as well as information exchange between different gazetteers. They ensured the quality of user-contributed information and improved querying and navigation using a semantics-based information retrieval approach. Smart et al. [46] presented a framework that accesses multiple gazetteers and digital maps in a mediation architecture for a meta-gazetteer service using similarity matching methods to conflate the multiple sources of place data in real time. Lieberman et al. [34] introduced a heuristic method to recognize toponyms and merging lists of them into *comma groups*. Toponyms in comma groups share a common geographic attribute and determine the correct interpretation of the place name. Zhang et al. [55] developed a supervised machine learning scheme to weigh the different features of a world gazetteer and fields of a Twitter message and to create a model that will prefer the correct gazetteer candidate to resolve the extracted expression. Middleton et al. [37] employed OpenStreetMap and used spatial filtering based on dynamically declared focus areas to generate inverted indexes for the geo-spatial entity recognition.

B. Language Models

The second class of geotagging approaches rely on the construction of large-scale geographical Language Models (LM) from geotagged corpora of text annotations, which act as *training sets* for the model. The goal of LM is to generate a probabilistic geographic model, which, given an arbitrary piece of text, produces probability estimates that the input text was generated (or originates) from specific locations across the globe. In a typical LM approach, a large corpus of geotagged text items is used for generating (training) the model. This typically takes the form of a set of geographic clusters (discrete areas) or a regular grid of cells covering the surface of the earth. Each such cluster or cell is associated with keyword frequency statistics that are used to generate location estimates for arbitrary pieces of text.

One of the earliest LM approaches was presented by Serdyukov et al. [43], where a predefined grid of cells is considered, and the prior probabilities for multimedia tags of a training corpus are computed based on the neighborhood of the cells where they appear. Hauff et al. [14] attempted to overcome the limitation of the fixed grid introducing disjoint dynamically sized cells. O'Hare and Murdock [39] proposed a statistical grid-based LM approach, which makes use of the Word-Document model, and they investigated several ways to estimate the models based on term and user frequency. Another approach was presented by Van Laere et al. [51], who first cluster the training corpus and then use the χ^2 feature selection criterion to create a vocabulary for every cluster. They also

[1] https://github.com/MKLab-ITI/multimedia-geotagging

[2] http://www.geonames.org/

[3] https://www.openstreetmap.org

[4] https://developer.yahoo.com/geo/geoplanet/

extended their approach in two ways: a) using the Dempster-Shafer theory of evidence to combine estimation from different granularities and to determine the most probable estimation [52]; b) using different term selection techniques, based on kernel density estimation and Ripley's K statistic to improve geotagging accuracy [50].

C. MediaEval Placing Task

1) Task description: MediaEval is an annual international benchmarking initiative that includes a number of multimedia analysis and retrieval tasks. Within its context, the Placing Task (PT) is dedicated to the geotagging problem using a corpus of geotagged Flickr images and videos for reference. Participants are challenged to estimate the locations (in terms of latitude and longitude) of items in a predefined test set using another set of items for training. Every year the released training and test sets are determined by the task organizers. In terms of evaluation, the submitted runs are benchmarked based on their *precision in different ranges* and their median distance error. The circular ranges vary from 1m to 1000km covering different geotagging granularities. The released datasets, the evaluation methods and the results of the participating approaches are further described in Section IV and are used as the state-of-the-art performance to compare with.

2) State-of-the-art geotagging systems: The participating systems in MediaEval PT over four years (2013-2016) are presented in Table I. Systems are classified depending on whether they use one or more of the popular geotagging approaches, namely: *Language Models* (LM), *Textual Analysis* (TA), *Visual Analysis* (VA), *Multimodal Fusion* (MF), *User Modelling* (UM) and *External Resources* (ER)[5]. Among the participating systems, the approach presented in this paper is an extension of the one originally tested in MediaEval PT 2014 [22] and then extended in PT 2015 [26] and PT 2016 [27].

TABLE I
PARTICIPATING SYSTEMS IN MEDIAEVAL PT CLASSIFIED BASED ON APPROACH.

Approach	LM	TA	VA	MF	UM	ER
Baseline [53]	✓					
Cao et al. [4] [2]	✓				✓	✓
Choi et al. [7]		✓	✓	✓		
Davies et al. [11]		✓	✓	✓		✓
Duong-Trung et al. [12]		✓				
Ferrés et al. [13]	✓	✓		✓		✓
Kelm et al. [19]		✓	✓	✓		✓
Kordopatis et al. [26] [27]	✓		✓	✓		✓
Kordopatis et al. [22]	✓		✓	✓		
Kordopatis et al. [25]		✓	✓	✓		
L. Li et al. [30] [31] [29]		✓	✓	✓		
X. Li et al. [33]			✓			
Muñoz et al. [38]		✓	✓	✓		
Popescu et al. [41] [40]	✓				✓	✓
Singh et al. [45]	✓		✓	✓		✓
Subramanian et al. [47]	✓					

III. PROPOSED APPROACH

The proposed approach relies on a LM that is built by calculating term occurrence probabilities from processing a massive amount of geotagged items of a training set D_{tr}. Given the generated model, it is then possible to estimate the location (in terms of latitude and longitude coordinates) for every query item m in a set of test items D_{ts}. In the case of Flickr images and videos (used as experimental test bed in this work), the metadata used include the tags, title, user id, image id and description. The metadata used to build the LM are the tags and titles of the items in D_{tr}. The initial LM is further refined through feature selection and weighting. Finally, the location estimation system employs two more steps (multiple grids, similarity search) oriented to achieve more accurate estimation in finer granularities. An overview of the proposed approach is illustrated in Figure 2. The system response can be modelled as a function $G_{pr}(m)$ that produces a location estimate for media item m. Given the ground truth location of item $G_{ref}(m)$, the geotagging precision of the system P at range R is computed based on Equation 1.

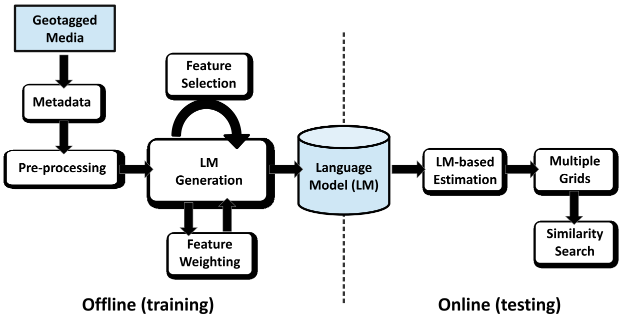


Fig. 2. Overview of proposed geotagging approach.

$$P@R = \frac{|\{m|d(G_{pr}(m), G_{ref}(m)) < R\}|}{|D_{ts}|} \tag{1}$$

where $d(x, y)$ is the geodesic distance between points x and y and $|D_{ts}|$ is the total number of items in test set. Also, the median distance error is computed, which is the median of the estimation errors across all test items in D_{ts}, i.e. the distances between the predicted and actual locations.

A. Pre-processing

A pre-processing step is first applied to determine the term set T_m of every item m. In the case of Flickr images and videos, their tags and titles are utilized to form the items' term sets. Initially, they are URL-decoded[6] and tokenized. All accents, punctuation and symbols are removed, all characters are transformed to lowercase, and all tokens consisting of numerics are removed. Additionally, there are multi-keyword tags, of which the keywords are linked by whitespaces. These multi-keyword tags are further split to single tags (e.g., `statue of liberty` is split into `statue`, `of` and `liberty`) and these are then added to the resulting tag set (if not already included). As a result, multi-keyword tags are included both as a whole and as separate tokens. The purpose of this operation is to increase the influence of multi-keyword tags on the geotagging

[5] These include gazetteers, online services such as translators and geocoders, geo-referenced collections, etc.

[6] This is specific to the MediaEval Placing Task dataset: texts in different languages are URL encoded.

results, and reduce the one of frequently occurring terms (e.g. `new`, `san`). After pre-processing, several items in D_{tr} are left with no tags and title and are hence disregarded from the remaining steps.

The terms of the resulting term sets associated with input items are considered as their representative features and are further processed for training the geotagging approach. The set of all unique terms of all items in D_{tr} is denoted as T.

Note that the same pre-processing is applied on the test items before the actual location estimation process, since the format of the test set is the same. Nevertheless, for estimating the location of a query item, its description is only used in cases where the term set from its tags and title is empty or did not generate any estimated location. Item descriptions are not used in any other case, since this would lead to less accurate results due to the fact that descriptions are sometimes very long and may refer to multiple locations, often irrelevant to the main location of the item. This was experimentally confirmed.

B. Language Model

The LM is constructed using a scheme that was originally presented in [40]. According to this, a rectangular grid C of cells is considered at granularity g and a map of term-cell probabilities is generated. Figure 3 illustrates an example of an LM cell with its term-cell probabilities. In the cell that lays upon the New York city, terms `nyc`, `manhattan`, `york`, etc. have high probability; in contrast, general interest terms (e.g. `new`) are assigned lower probability scores, because they are commonly used in many other cells around the globe. Note that the particular example is just for visualization purposes and the illustrated grid does not accurately reflect reality. Since the earth is ellipsoid and its projection on a 2D plane causes deformation, cells become shorter as they approach the poles instead of having the same side height across the entire globe.

For the needs of the proposed approach, four grids at different granularities are considered. Starting from coarser to finer granularity, we consider grid cells at the level of `region`, `city`, `neighborhood` and `street` with sides of 1°, 0.1°, 0.01° and 0.001° for both latitude and longitude, corresponding to geodesic distances of approximately 100km, 10km, 1km and 100m near the equator, respectively. The default LM for our system is built at a `neighborhood` level, since this was empirically found to lead to better results. The remaining grid levels are defined for the following reasons: a) to support the formulation of the multiple grids technique (cf. section III-E); b) to support the generation of geotagging models that are better tuned for other granularities (e.g. to produce coarser or much more detailed location predictions).

The main purpose of the LM is to estimate the most likely cell $c \in C$ for a query item m based on its term set T_m. The probability p of a term t in a particular cell c is calculated as the total number of different Flickr users that used t inside c, divided by the total count of users over the entire grid C. For simplicity, the total count of different users over the whole grid C that used a specific term t will be referred as the user count of t. Eventually, the term-cell probability $p(t|c)$ is calculated for every term $t \in T$ according to Equation 2.

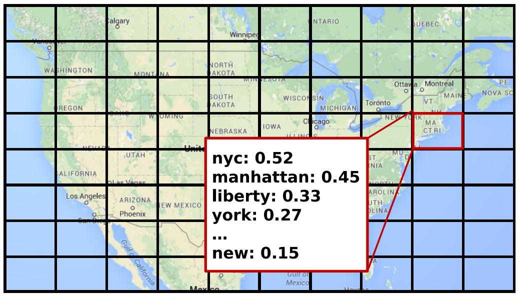


Fig. 3. Visual example of LM term-cell probabilities.

$$p(t|c) = \frac{N_{u,c}}{N_t} \tag{2}$$

where $N_{u,c}$ is the number of users in D_{tr} that used the term t inside the borders of cell c, and N_t is the user count of term t in all cells. Note that a user can be counted in N_t more than once. If a user u is found in multiple cells, every time he/she is found in a different cell, he/she is considered as a new user and increases the total count of users.

To assign a query item to a cell, the probability of each cell of C is first calculated summing up the contributions of each term in T. Then, the cell with the highest probability is selected as the *most likely cell* (*mlc*) according to Equation 3.

$$mlc_m = \underset{c_i \in C}{\operatorname{argmax}} \sum_{k=1}^{|T_m|} p(t_k|c_i) \tag{3}$$

where T_m is the set of terms for item m, and $p(t_k|c_i)$ is the term-cell probability for term $t_k \in T_m$ in cell $c_i \in C$. As a result, the centroid of the estimated mlc may be considered as a coarse location estimation for query item m and is denoted by $G_{pr}^{lm}(m)$. If during this process there is no outcome (i.e. the probability for all cells is zero), then the description of the query item m (in case of Flickr images and videos) is utilized. For items where there is no result (e.g. images completely lacking text annotations), their location is set equal to the centre of the most populated cell, which is a kind of maximum likelihood estimation.

C. Feature selection

To increase the robustness of the model and reduce its size, a feature selection scheme is necessary, which reduces the set of features (terms) into a compact set $T_s \subset T$, with $|T_s| << |T|$. The goal of this refinement is the selection of the most appropriate terms based on their ability to discriminate different locations from each other. A first filtering step removes all terms in T that are used by only one user, as they are considered user- and dataset-specific. The reduction of the term set through this simple step is dramatic (up to 87%). The remaining terms are then ranked and filtered on the basis of three measures: *accuracy*, *spatial entropy* and *locality*.

1) Accuracy: This was originally proposed in [23] as a means of quantifying the geotagging capability of terms and correlating their occurrence with correct location estimates. To calculate the accuracy of a term $t \in T$, a scheme similar to cross-validation is employed. First, D_{tr} is partitioned into q folds. The number of partitions is empirically selected; in our experiments, it is set to 10. Subsequently, one partition D_{tr}^{p} at a time is withheld, and the remaining $p-1$ partitions are used to build the LM. Using this LM, the location of every item in the withheld partition is predicted as described in subsection III-B. Then, accuracy is computed as the ratio of the correctly geotagged items where the term appears over the total number of items where the term appears (Equation 4).

$$\alpha(t) = \frac{|\{m|t \in T_m \wedge d(G_{pr}(m), G_{ref}(m)) < R\}|}{m_t} \in [0, 1] \quad (4)$$

where $\alpha(t)$ is the accuracy of term t, the numerator determines the total number of correctly geotagged items (within range R) associated with t in D_{tr}^{p}, and m_t is the total number of items in D_{tr}^{p} where the term t occurs. The selected range R is considered as a system hyperparameter and its effect is explored in Section IV-A. The grid used for the accuracy calculation is always the same as the grid of the default LM.

To perform feature selection, terms in T are sorted in descending order based on their accuracy generating a ranked set of terms T_{α}. Terms with the same accuracy score are further sorted based on their frequency. In that way, it is possible to build an LM with a target number of terms N by selecting the first N elements of T_{α}.

2) Spatial Entropy: This feature selection measure attempts to capture the *spatial ambiguity* of terms [22]. The measure is computed by quantifying the stochasticity (or randomness) in the spatial distribution of the term. To this end, the spatial entropy of a term is computed based on the Shannon entropy formula on the term-cell probability distribution (Equation 5).

$$se(t) = -\sum_{i=1}^{|C|} p(t|c_i) \log p(t|c_i) \quad (5)$$

where $se(t)$ is the spatial entropy of term t and $p(t|c_i)$ is the term-cell probability of t in cell $c_i \in C$. In a sense, spatial entropy expresses the amount of information conveyed by term t regarding a cell c. Terms appearing in few cells tend to have low spatial entropy values (high information), while terms with a relatively uniform distribution over many cells have high entropy values (low information).

For feature selection, the terms in T are sorted in ascending order based on their spatial entropy resulting in a ranked term set denoted as T_{se}. The grid granularity that is used for the calculation of spatial entropy is considered as a system hyperparameter and its effect is explored in Section IV-A. Hence, the LM can be built based on a target number of N terms by selecting the first N elements of T_{se}.

3) Locality: In [50], Van Laere et al. introduced two approaches to capture the spatial discrimination of a tag: a method based on Kernel Density Estimation (KDE) [44], and one based on Ripley's K statistic [42]. However, both methods are computationally expensive. Thus, the main objective of defining locality has been to come up with a measure that is equally discriminating as the Ripley's K statistic, but computationally much lighter so that it is possible to compute over massive datasets.

The computation of locality is based on the number of different users that make use of a term in the same cells. The locality score of a term is calculated based on the term user count and the *unique users* that have used it in a cell of the grid. The users that have used a term t in a cell c are assigned to the unique user set $U_{t,c}$ of that particular cell. Every unique user $u \in U_{t,c}$ is associated with all other users included in the set, i.e. with $|U_{t,c}|-1$ other unique users in c. Locality derives from the summation over all such user associations across all grid cells divided by the total user count. As a result, cells with only a single user in their set do not affect the locality calculation. Locality is computed according to Equation 6:

$$l(t) = \frac{\sum_{c \in C} \sum_{u \in U_{t,c}} |U_{t,c}| - 1}{N_t} = \frac{\sum_{c \in C} |U_{t,c}|(|U_{t,c}| - 1)}{N_t} \quad (6)$$

where $l(t)$ is the locality score of term t, N_t is the user count of t, C denotes the set of cells and $U_{t,c}$ denotes the set of users that used tag t in cell c.

Similar to the previous feature selection measures, the terms in set T are sorted in descending order based on locality. Terms with the same locality score are further sorted based on their user count. In that way, a ranked term set T_l is generated. The grid granularity that is used for the calculation of locality is considered as a system hyperparameter and its effect is further explored in Section IV-A. Similar to the aformentioned measures, it is possible to build an LM with a target number of terms N by selecting the first N elements of T_l.

D. Feature Weighting

Feature weighting aims to make geotagging more accurate by giving more importance to terms with good accuracy, spatial entropy and locality scores. To this end, weighting scores w_{α}, w_{se} and w_l are computed for each term in the set of selected features T_s based on scores α, se_s and l respectively.

Since accuracy scores are already in the range [0,1], the generated weights are equal to them, i.e. $w_{\alpha} = \alpha$.

The computation of spatial entropy weights is a bit more complicated. Based on our observations, terms with either too high or too low entropy values typically carry no geographic information. For instance, terms with too low entropy values tend to be user-specific. In contrast, very high entropy values indicate terms that are widely spread across the globe. Such terms carry no geographical interest (e.g., `baby` and `fun`) and therefore their influence on location estimation needs to be suppressed. To this end, after experimenting with different distribution functions that appeared to fit the empirical spatial entropy distribution, we selected the gamma distribution (Equation 7) for transforming the spatial entropy values.

$$F(se(t)|a, b) = \frac{1}{b^a \Gamma(a)} se(t)^{a-1} e^{\frac{-se(t)}{b}} \quad (7)$$

where F is the probability density function of the gamma distribution, and parameters a, b are the shape and scale parameter respectively, which are learned based on the empirical

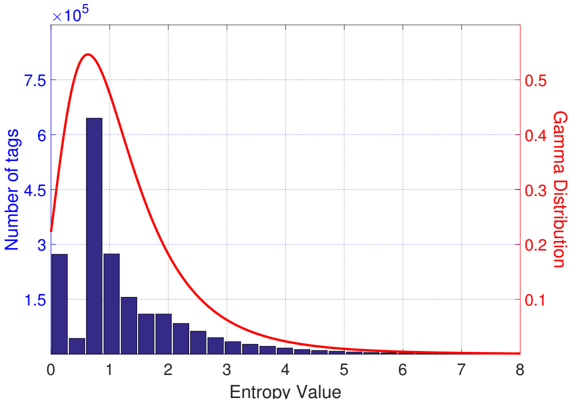


Fig. 4. Histogram of spatial entropy values based on `city` level grid and the corresponding probability density function of the fitted gamma distribution.

spatial entropy distribution on D_{tr}. The transformed values are then used as the spatial entropy weights. Figure 4 illustrates the distribution of entropy values along with the fitted density function. These weights are then normalized (by dividing them with the maximum value in the new distribution) to bring them in the range $(0, 1]$.

Locality scores are quite sensitive to the respective term frequencies. To mitigate this sensitivity, terms in T_l (set of selected terms ordered by locality) are assigned weights proportional to their position in T_l (Equation 8).

$$w_l = \frac{|T_l| - (j-1)}{|T_l|} \tag{8}$$

where w_l is the weight value of term t on the j-th position in the ordered set T_l. This weighting approach returns values in the range $(0, 1]$.

To combine the three weights, a linear scheme is applied:

$$\begin{aligned} w_t &= \omega_\alpha \cdot w_\alpha + \omega_{se} \cdot w_{se} + \omega_l \cdot w_l \\ \omega_\alpha + \omega_{se} + \omega_l &= 1 \end{aligned} \tag{9}$$

where w_t is the final term weight, w_a, w_{se} and w_l are the weights derived based on accuracy, spatial entropy and locality, respectively, and ω_α, ω_{se} and ω_l are constants that determine the effect of each weight, and summing to 1. The choice of each ω will be discussed in the evaluation Section IV.

Finally, after the computation of the final term weights, the estimation of the most likely cell for a query item m is performed using Equation 10.

$$mlc_m = \underset{c_i \in C}{\operatorname{argmax}} \sum_{k=1}^{|T_m|} w_{t_k} p(t_k | c_i) \tag{10}$$

Table II presents the top 15 terms based on the generated weighting scores (computed at the `neighborhood` grid). Terms ranked by accuracy correspond to very specific locations or events (Table II(a)), those ranked by the spatial entropy weight correspond to landmarks or points of interest, while locality ranked terms correspond to cities. This means that the three weighting scores capture different types of geographic information and are all valuable in deriving a total weighting score that captures the importance of a term for location estimation.

TABLE II
Top 15 terms based on their weights.

(a) accuracy	(b) spatial-entropy	(c) locality
barclays center arena	kaiyukan	london
romanische kunst	cancale	paris
untermyer	collins street	nyc
passim	gatorland	eiffel
rigid inflatable boats	friedrichstraße	san francisco
cpbrasil	how long is now	barcelona
national police week	fairford	york
festineuch	beaumaris	francisco
lincoln imp	plaza del pilar	berlin
david bell	amarapura	louvre
gangale	queens house	manhattan
frannie garretson	stintino	amsterdam
protest photography	marischal	rome
barnsdall art park	port macquarie	brooklyn
buskersbern	roosevelt island	new york

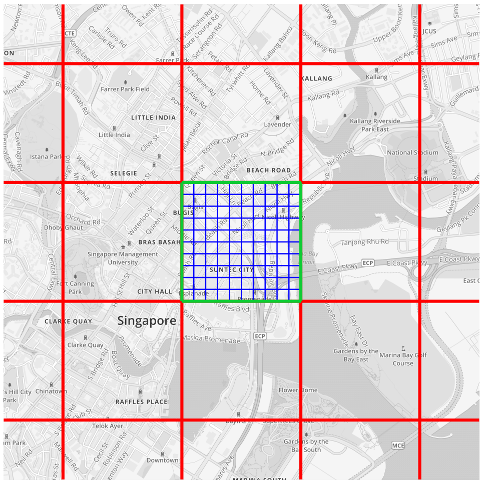


Fig. 5. Depiction of the Multiple Grids technique using Singapore as example.

E. Multiple Grids

To ensure more reliable and accurate estimations, the estimations based on two LMs at different granularities are combined into a single estimation. The most likely cells for the coarser and finer granularities are denoted as mlc^c and mlc^f respectively. The combination of the two is performed using the *multiple grids* method originally proposed in [22]: if the estimated fine granularity cell mlc^f falls within the borders of the estimated coarse granularity cell mlc^c, then the prediction of the finer granularity is considered reliable and the final estimated cell is mlc^f. Otherwise, the final estimated cell is mlc^c, since coarser granularity LMs are considered more reliable by default, assuming that more data per cell are used for their creation, and hence the resulting probabilistic analysis is more robust. The process is outlined by the Equation 11 and illustrated in Figure 5.

$$mlc^{mg} = \begin{cases} mlc^f & \text{if } mlc^f \subseteq mlc^c \\ mlc^c & \text{otherwise} \end{cases} \tag{11}$$

where mlc^{mg} is the estimated cell from the multiple grids technique.

The red granularity grid in Figure 5 represents the coarse granularity and the blue the fine one respectively. The most likely cell of the red grid mlc^c has been colored green. If the most likely cell of the finer grid mlc^f is one of the blue cells of the particular example, then the final estimation of the system would be based on mlc^f; otherwise, in case that mlc^f is outside mlc^c, then the whole green cell mlc^c will be used to produce the final estimation.

A related but much more complicated method was used in [52], where the authors combined the results of multiple geographic models using the Dempster-Shafer theory of evidence to determine the most reliable prediction. However, the approach in [52] is not based on regular grids, but on spatial clusterings of items using k-means for different values of k.

F. Similarity Search

Given the most likely cell for a query item, a further refinement is conducted using the *similarity search* technique of [51]. This is done by identifying the k most similar items to a query item from D_{tr} in terms of textual similarity, and combining their locations (weighted by their similarity to the query). To this end, we first compute the textual similarity between the query item m and every item in D_{tr} that falls inside the borders of mlc^{mg} by use of the Jaccard similarity on the corresponding sets of terms (Equation 12).

$$J(T_m, T_i) = \frac{|T_m \cap T_i|}{|T_m \cup T_i|}, m \in D_{ts}, c_i = mlc^{mg} \tag{12}$$

where T_m, T_i denote the term sets of items m and i, respectively, and c_i is the cell of item i.

After calculating the similarity with every item in the mlc, the top k most similar items to the query are selected and the final estimation is the centre-of-gravity of their locations, weighted by the similarity values. The estimated location for item m is determined by Equation 13.

$$G_{pr}^{ss}(m) = \frac{1}{k}\sum_{i=1}^{k} J(T_m, T_i)^a \cdot loc(i) \tag{13}$$

where parameter $\alpha \in [0, +\infty)$ determines how strongly the result is influenced by the most similar items and $loc(i)$ denotes the vector of coordinates for item i. For the accurate calculation of the average location [51][7], the location coordinates of the k items are first transformed to the Cartesian (x, y, z) system and after the computation are transformed back to spherical coordinates (latitude, longitude).

IV. Evaluation

For the evaluation of the proposed approach, we use the precision P in various ranges R ($P@R$), computed by Equation 1, and the median distance error, which is the median of the estimation errors across all test items in D_{ts} in terms of the distance between the predicted and the actual location.

The datasets used for building the LMs and testing the approach are all derived from YFCC100M [48]: D_{tr} consists of all the images and videos in YFCC100M that are geotagged, excluding all items of users that are also included in the test set to avoid over-fitting and providing misleading results. D_{ts} was released by the organizers of MediaEval PT, and its different versions are presented in Table III. All datasets used in this section consist of images and videos (except for the 2013 version that contains only images). The way that both videos and images are processed is identical.

TABLE III
Four editions of MediaEval PT used for testing.

Year	Training Set		Test Set		Origin
	images	videos	images	videos	
2013 [15]	8,539,050	-	262,000	-	Flickr
2014 [8]	5,000,000	25,000	500,000	10,000	YFCC100M
2015 [5]	4,672,382	22,767	931,573	18,316	YFCC100M
2016 [6]	4,991,679	24,955	1,497,464	29,934	YFCC100M

The calculation of geodesic distance between the estimated and the real location of an item is based on Karney's algorithm [18][8], which relies on the assumption that the shape of the earth is an oblate spheroid. This algorithm produces more accurate distances than methods such as the great-circle distance that assume the shape of the earth is spherical.

For the sake of brevity, we use the following short names for the components of the approach: LM (Language Model), FS_x (Feature Selection), FW (Feature Weighting), MG (Multiple Grid), and SS (Similarity Search).

A. Fine Tuning

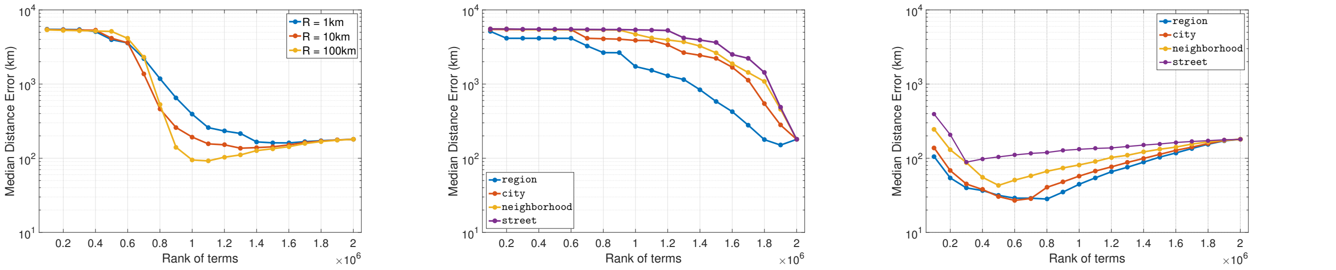
A set of 100k items was withheld from the training set and was used for experimenting with different values of the method's parameters to optimize performance in terms of $P@1km$ and median distance error. In particular, parameter tuning was carried out with respect to the thresholding used for each feature selection measure, and the ω factors used for feature weighting in Equation 9.

1) Feature Selection: In this paragraph, we evaluate the performance of an LM built at `neighborhood` level, involving a certain number of terms denoted by T_α, T_{se} or T_l, as selected by the three feature selection measures of section III-C. After the initial filtering, where the terms used by a single user were removed, the total amount of remaining terms is ≈2M. The objective of this step is to minimize the median error of the location estimations by selecting a certain number of top terms from the ranked sets T_α, T_{se}, T_l. As described in section III-C, for tuning term selection with respect to accuracy, different values of parameter R are considered (i.e. 1km, 10km and 100km); instead, for tuning with respect to spatial entropy and locality, the four granularity levels are compared.

Figure 6 depicts the performance of the different feature selection measures in terms of median error. The locality measure appears to lead to the best results. The minimum median error of 28km is reached when the top 600k tags are selected at the `city` grid (Figure 6(c)). Regarding accuracy, the best performance is achieved for $R = 100km$ at 1.1M tags with 92km median error (Figure 6(a)). Finally, spatial entropy performs the worst in comparison to the other two measures

[7] http://www.geomidpoint.com/calculation.html

[8] http://geographiclib.sourceforge.net/geod.html

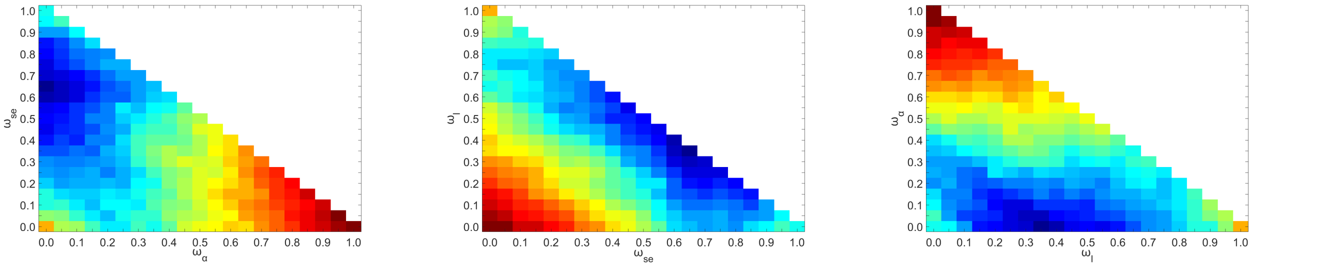


(a) accuracy

(b) spatial entropy

(c) locality

Fig. 6. Median distance error (km) of the approach when selecting an increasing number of features (terms) ranked based on accuracy, spatial entropy and locality scores. Accuracy metric is tuned based on the range R, spatial entropy and locality are tuned based on the granularity grid used.



(a) ω_α - ω_{se}

(b) ω_{se} - ω_l

(c) ω_l - ω_α

Fig. 7. Geotagging precision at 1km (%) of the approach using the different values of ω_α, ω_{se}, ω_l. Blue color indicates higher precision, whereas red color lower precision.

for all grid levels, since it needs almost all features to perform equally well as other much more succinct models produced using locality and accuracy. In particular, the `region` grid performs the best with 150km median error when the top 1.9M tags are selected (Figure 6(b)).

The final term set used to build the LM is the intersection of sets T_α, T_{se} or T_l that maximizes geotagging performance in terms of median error, i.e. $T_s = T_\alpha^{max} \cap T_{se}^{max} \cap T_l^{max}$, where T_s is the selected term set and T_x^{max} is the subset of T_x that minimizes the median error. The feature selection scheme is initially applied to significantly reduce the amount of processed terms and minimize the requirements of the approach in terms of computation time and storage space. The total size of the selected term set T_s is 550,050 terms, which equals to approximately 4% of the initial term set.

2) Feature Weighting: The goal of this step is to determine the combination of weight parameters that maximize geotagging performace. The performance metric used for the tuning is $P@1km$, because it was found to be sensitive to small variations between the setups and hence can express the differences in performance more accurately compared to median error. The results of the approach for different values of ω_α, ω_{se}, ω_l are illustrated in Figure 7. In each plot, the dependence of performance on every pair of ω parameters is presented; at each point, the third weight value derives from the constraint of Equation 9. The triangular form is due to the fact that the sum of three parameters can never exceed 1. Blue color corresponds to parameter values leading to higher $P@1km$, while red to parameter values leading to lower $P@1km$. The plots indicate that higher locality and spatial entropy weights lead to better geotagging performance. In the first parameter pair, deep blue is concentrated at the left side. In this area ω_α is equal to zero and ω_l has values 0.3-0.5 indicating that ω_{se} has values 0.5-0.7. This parameter set is further supported by the rest of the plots. In the ω_l-ω_{se} pair, the deepest blue color is in the hypotenuse of the triangle and in the ω_α-ω_{se}, it is at the bottom side of the triangle. As a result, the values of ω_α, ω_{se}, ω_l are selected to be equal to 0.0, 0.65, 0.35 respectively.

It is noteworthy that ω_α is set to zero, even though accuracy was found to be beneficial for feature selection. In contrast, ω_{se} was found to have the highest weight among others, despite the fact that spatial entropy scoring had only minimal impact on the feature selection process. This observation demonstrates the complementarity among the three feature selection and weighting measures and the fact that all three of them contribute to optimizing the geotagging performance of the approach.

B. Performance Analysis

This section explores in detail the performance of the approach. All experiments use the parameter set that was selected in the previous section. The objective of the discussion is to highlight the contribution of each of the processing steps described in section III to the overall geotagging performance. The training dataset used for these experiments is the entire YFCC100M, excluding all items from users also appearing in the test set. Hence, the total number of items used for training is ≈40M. The test set used in all experiments is the one

TABLE IV

Geotagging precision (%) for five ranges and median geotagging error (km) for different configurations of the proposed approach.

LM	FS	FW	MG	SS	P@10m	P@100m	P@1km	P@10km	P@100km	m. error
✓					0.02	0.75	23.83	40.67	47.96	173
✓	✓				0.02	0.81	26.57	47.08	55.08	20
✓	✓	✓			0.02	0.81	26.83	47.85	56.06	16
✓	✓	✓	✓		0.18	7.15	27.18	47.85	56.06	16
✓	✓	✓	✓	✓	0.70	7.52	27.40	47.86	56.06	16

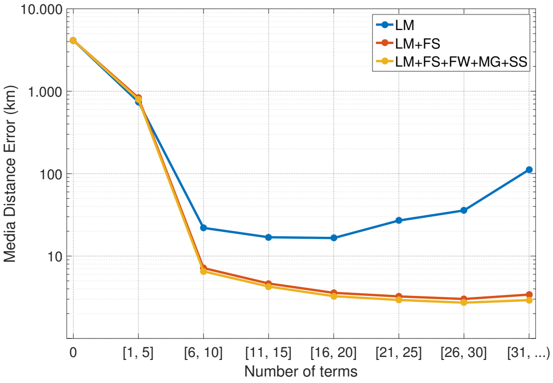


Fig. 8. Median geotagging error (km) with respect to the number of terms per item for different approach variations.

released by the organizers of MediaEval 2016 PT, comprising 1,527,398 items.

Table IV presents the performance of the framework when different processing steps are included. The base approach using solely the LM performs poorly in low ranges, e.g. it achieves $P@100m$=0.75%, and has a high median error of 173km. Applying the feature selection scheme (FS) described in section IV-A, the median error is dramatically reduced to 20km and the precision in medium/high ranges improves in absolute terms by more than 10%. Applying feature weighting (FW) further reduces the median error to 16km and leads to slight improvements in the medium/high ranges (10, 100km). The introduction of multiple grids (MG), leads to significant improvements (almost 10-fold) in the low ranges (10, 100m), e.g. reaching $P@100m$=7.15%. Finally, the best results are achieved by also integrating the similarity search step (SS), which leads to a performance of $P@100m$=7.52%, $P@1km$=27.40% and 16km median distance error.

Figure 8 depicts the performance of the approach on different subsets of images that differ with respect to the number of terms associated with them. It becomes obvious that the introduction of feature selection (FS) leads to dramatic improvements with respect to median error: while in the original LM-based approach, geotagging performance starts deteriorating for items with more than 15 terms, the FS-powered version of the system manages to retain stable performance for almost all items with more than 15 terms. As a result, the LM+FS version of the system achieves a median error of just 3.6km for items with a number of terms in the range [21,25], while applying the additional refinements (FW+MG+SS) leads to a further decrease of the error to just 3km for items with number of terms in the range [26,30].

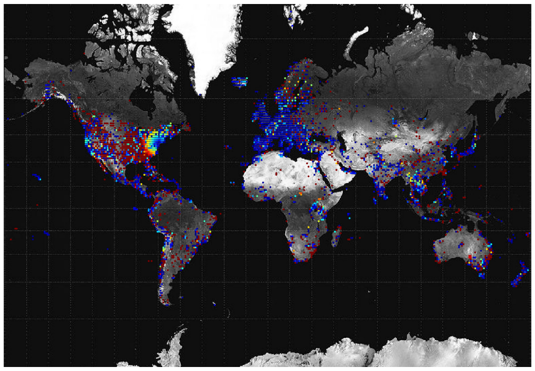


Fig. 9. Spatial distribution of median distance error per cell of the best system configuration. Deep blue color indicates median error < 100km, whereas brown red color indicates median error > 1000km.

Figure 9 illustrates the median geotagging error per cell across the globe. The cells with median error less than 100km are displayed with deep blue color, whereas those with more than 1000km are displayed with brown red color. It is noteworthy that in North America and Australia the dominant color is brown (very high error), despite the availability of much more training data and the prevalence of English text (which is expected to be easier to handle). In contrast, in Europe large areas are painted in blue, so in these areas the system worked considerably better. A possible explanation for the high error levels in the US and Australia is the potential ambiguity in town and city names (e.g., many American and Australian towns are named after popular European ones).

To further delve into this performance aspect, we compute the spatial entropy of terms in the test set based on the scheme presented in section III-C2 (at the `city` granularity level) and create a scatter plot of the terms' geotagging precision at 10km range in relation to their spatial entropy. The precision of a term is computed over the set of items that are associated with this term. The scatter plot, which is illustrated in Figure 10, comprises all terms that occur more than 100 times in the test set and are associated with at least two different places from the Geonames dataset[9] (in particular, the cities with a population above 1,000).

The plot reveals that terms with relatively large spatial entropy values tend to be associated with low precision scores, and vice versa. To further study the hypothesis that text annotations with ambiguous names are harder to geotag, we compute the median of the spatial entropy values of ambiguous terms with more than 100 occurrences (M_a=3.626) and split this set in two groups, one with terms that have spatial entropy less than M_a and one with the remaining ones. Accordingly, we generate two sets of items from the Mediaeval 2016 PT test set, denoting them as *low-ambiguity set* (`Low-AS`) and *high-ambiguity set* (`High-AS`) respectively. The geotagging performance for these two sets is reported in Table V. The results confirm our hypothesis that text annotations of low ambiguity can be geotagged with higher precision compared to those of high ambiguity. For instance, the low-ambiguity items were geotagged with a $P@100m$=15.22% and median distance

[9] http://www.geonames.org/

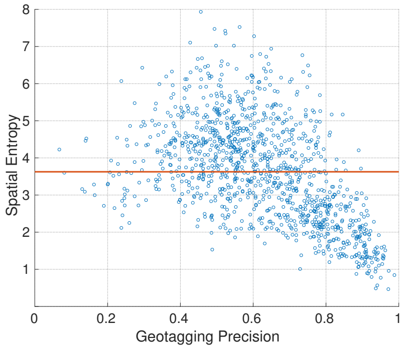


Fig. 10. Scatter plot of geotagging precision ($P@10km$) in relation to spatial entropy for the most frequently appearing ambiguous terms. The red line indicates the median of the spatial entropy values.

error of 1.09km. The accuracy for the highly ambiguous set is considerably lower, but still higher than the one obtained using the overall test set. This is attributed to the fact that the overall test set contains a considerable number of items with no or very little annotations, which is even more challenging for text-based approaches compared to ambiguous annotations.

TABLE V
GEOTAGGING PRECISION (%) FOR FIVE RANGES AND MEDIAN GEOTAGGING ERROR (KM) OF THE PROPOSED APPROACH ON LOW- AND HIGH-AMBIGUITY OF ITEMS.

	P@10m	**P@100m**	**P@1km**	**P@10km**	**P@100km**	**m. error**
`Low-AS`	1.36	15.22	48.55	78.88	85.63	1.09
`High-AS`	0.71	7.69	28.88	52.57	64.48	7.17

Finally, we benchmarked the geotagging performance of the proposed approach using either the locality or Ripley's K statistic for feature selection and weighting. The dataset used for training and testing is the MediaEval PT 2016 edition. The locality was calculated at the `city` level, while for computing Ripley's K a sample of 10,000 items per term was used. For feature selection, the terms with either locality or Ripley's K statistic greater than zero were selected. For feature weighting, the same process described at Section III-D was applied in both cases.

Table VI displays the results of the LM using either locality or Ripley's K statistic. Performance is very similar for both measures, with Ripley's K statistic performing slightly better. Yet, the proposed locality measure comes with a huge computational benefit: computing Ripley's K statistic required 465 min, whereas locality took just 7 min, i.e. it was approximately 66× faster and was calculated without any sampling. The sets of terms selected by the two measures (terms with value greater than zero) amount to 278,240 for locality and 304,419 for Ripley's K. The Jaccard similarity of the two sets is ≈89% meaning that both measures lead to highly similar sets (274,359 of the 278,240 terms selected based on locality had been also selected based on Ripley's K statistic).

A more nuanced performance analysis of the approach is presented in [24], which reveals how the performance changes depending on particular traits of the test dataset (e.g. when the dataset consists of images from a specific location, depict specific objects, etc.).

TABLE VI
GEOTAGGING PRECISION (%) FOR THREE RANGES AND MEDIAN GEOTAGGING ERROR (KM) OF THE LM USING FOR FEATURE SELECTION AND WEIGHTING EITHER LOCALITY OR RIPLEY'S K STATISTC.

	P@1km	**P@10km**	**P@100km**	**m. error**
LM+FS+FW (locality)	23.94	43.44	51.48	60
LM+FS+FW (Ripley's K)	24.02	43.72	52.02	56

C. *Comparison against geoparsing methods*

In this section, the performance of the proposed approach is compared with two state-of-the-art geoparsing approaches on the MediaEval 2016 PT dataset. The results for the proposed approach were obtained using the LM with all the refinements and using the entire YFCC100M for training (after excluding all items from users also appearing in the test set). The selected geoparsing approaches are the following:

- Zhang et al. [55]: This is a preference learning approach that, given an input text, detects the GeoNames entities (if any) that are mentioned; the approach makes use gazetteer-based features and a corpus of geotagged tweets in order to build an accurate geoparsing model.
- DBpedia geoparsing [28]: This is a simple geoparsing scheme that provides the input text as a geo-query to DBpedia, i.e. a query that is limited only to objects associated with geographical information, and uses the returned DBpedia entities as location candidates.
- CLAVIN[10]: This is a widely used open-source geoparsing library; however, it produced location estimates for only a tiny fraction of the test items and therefore led to very poor results. For that reason, its results are not included in our comparison.

The same testing protocol was used as in the case of our proposed approach: The tags and titles of the test items were fed as input to the geoparsing methods. In cases where no estimation was returned, the item descriptions were then provided. The only text pre-processing step on the input data is URL-decoding. Since geoparsing methods produce a set of geographical entities (each of which associated with a pair of lat/lon coordinates) and not necessarily a single location, it was necessary to devise a location selection step to determine a single location per input item. To this end, two variations were considered:

- **optimal**: The distance between all candidate locations and the ground truth location is computed, and then the location with the smallest distance is selected. This corresponds to an upper bound of performance (i.e. case where the candidate selection step would be perfect).
- **random**: According to this, a random location is selected for every item among the candidate locations. This process is repeated 10 times and the mean performance is reported. This correspond to an average estimate of performance for such methods.

Table VII presents the results of the comparison. The proposed approach outperforms both geoparsing methods at a very large margin, even in their optimal variation. In all ranges, especially low ones (i.e. 100m, 1km), precision

[10] https://github.com/Berico-Technologies/CLAVIN

TABLE VII
GEOTAGGING PRECISION (%) FOR FOUR RANGES AND MEDIAN GEOTAGGING ERROR (KM) OF THE PROPOSED APPROACH AND THE VARIANTS OF THE COMPETING GEOPARSING METHODS.

	P@100m	P@1km	P@10km	P@100km	m. error
Zhang (optimal) [55]	1.77	13.71	37.04	48.68	131
Zhang (random) [55]	0.68	5.78	17.65	27.87	1148
DBpedia (optimal) [10]	1.78	10.94	29.49	37.88	891
DBpedia (random) [10]	1.31	8.74	25.22	34.05	1151
proposed approach	**7.52**	**27.40**	**47.86**	**56.06**	**15**

TABLE VIII
GEOTAGGING PRECISION (%) FOR FIVE RANGES AND MEDIAN GEOTAGGING ERROR (KM) OF THE TEXT RUNS `RUN1` FOR PARTICIPANTS IN THE MEDIAEVAL PT 2013 AND FOR THE PROPOSED APPROACH.

	P@100m	P@1km	P@10km	P@100km	m. error
`CERTH` [25]	2.96	10.26	23.52	36.25	651
`UoS` [11]	5.43	23.15	37.70	43.83	451
`SCUT` [2]	4.90	20.74	42.95	**55.26**	**38**
`CEA LIST` [40]	**7.41**	**26.00**	42.77	50.03	99
`VIT` [47]	0.06	0.74	3.92	15.24	6183
`RECOD` [29]	6.07	20.13	37.60	47.64	168
proposed approach	6.23	25.21	**44.02**	53.25	47
p. approach (YFCC)	7.74	26.71	44.94	54.78	32

scores are significantly better. Similar results are also obtained when comparing the respective median errors: the proposed approach achieves > 8× lower error than the second best.

D. MediaEval Placing Task

In this section, the proposed system is benchmarked against the participating teams in the MediaEval PT of years 2016 [6], 2015 [5], 2014 [8] and 2013 [15]. Every year, participants were asked by the organizers to submit their approaches using the released dataset. However, the volume and origin of the released datasets vary from year to year. Table III presents those details for every edition of the task.

The proposed system is tested with every one of the four datasets using the corresponding evaluation setup, and its results are compared with the reported results of the approaches that participated in the respective edition. For years 2014, 2015 and 2016, the test set of each year is a superset of the one used in the previous editions, i.e. 2015 test set is a superset of the 2014 set, and 2016 test set a superset both 2015 and 2014 sets. Consequently, their results are presented in a single concatenated table to facilitate comparison, and provide a comprehensive view into the performance of all methods. The detailed method results were provided by the MediaEval PT organizers. In all cases, the instance of the proposed approach was built based on the tuning process of section IV-A. In addition to presenting the results when training the approach with the training sets provided by the organizers each year, we also report the performance of the method (under the entry *proposed approach (YFCC)*) when it is trained with a much larger set, i.e. the full YFCC100M (after removing items of which the owners appear in the test set).

1) MediaEval 2013: The proposed approach was evaluated based on the data of MediaEval 2013 Placing Task, i.e. a training set of ≈8.5M items and a test set of 262k items.

The results of the participating approaches are illustrated in Table VIII along with those of the proposed approach. In this case, `run1` was not strictly restricted to the use of text-only information; to make comparison fair, the table includes only the results from teams that submitted a text-only run. The proposed approach ranks firmly between the first and second place in the different precision ranges: it achieves 25.21% at $P@1km$ and the second best median error (47km). Overall, the proposed scheme is highly competitive and outperforms most of the participating systems.

2) MediaEval 2014, 2015, 2016: The comparative results of the proposed approach for each of the three editions of PT (2014-2016) are presented in Table IX. Three versions of the proposed approach are presented given that each year a different training set was released. Note that earlier versions of the proposed approach had contested under the names `SocialSensor` [22] and `CERTH/CEA LIST` [26] [27].

The proposed approach ranks in the first or second place in all years and performance measures; in particular, it achieves significantly better results in precision ranges $P@1km$ and $P@10km$ ($P@1km$ = 25.45% and $P@10km$ = 45.76% on the 2014 test set), as well as the best median error (25km on the 2014 test set). In addition to achieving top results against all state-of-the-art approaches, it is noteworthy that the performance of the proposed approach is very stable across the various datasets, exhibiting only minimal variance with respect to the different training and test sets.

Also, as expected, the performance of the approach is further improved when it is trained on the much larger set. For instance, in the case of MediaEval PT 2013, we achieved an absolute increase between 1.5% and 1.9% in terms of $P@R$ and a reduction of 15km in median error. Similar gains are also achieved for the 2014-2016 editions of the task, with a notable absolute increase of ≈ 2.5% in $P@1km$ and $P@10km$ for all test sets. This observation points to an interesting trade-off and question of whether a measurable but moderate increase in the geotagging performance justifies the investment in a very large increase (almost 10-fold) in the training set.

E. Discussion: Geotagging on other datasets

Through our experimental study, we presented compelling evidence that the proposed approach achieved excellent geotagging accuracy, outperforming all text-based approaches that have competed in MediaEval Placing Task, as well as a couple of popular geoparsing approaches. However, we need to recognize that all tests have been carried out on Flickr collections. Hence, it should not come as a surprise that applying the proposed approach on arbitrary (non-Flickr) text data, such as tweets and news articles, may lead to suboptimal or even unsatisfactory performance. The main reason for such an expectation stems from the fact that the underlying LM was built using a dataset of Flickr image and video annotations. Even though the used training sets were very large scale, one needs to bear in mind that they exhibit specific characteristics that are tightly associated with the "typical" content that is published on Flickr (ranging from photos of touristic sites, events, social activities, artistic creations, etc.).

It is worth mentioning that we have carried out experiments on Twitter data with encouraging results, even though the underlying LMs were the same as the ones that we used for

TABLE IX

GEOTAGGING PRECISION (%) FOR THREE RANGES AND MEDIAN GEOTAGGING ERROR (KM) OF TEXT RUNS (RUN1) FOR PARTICIPANTS OF EACH YEAR OF MEDIAEVAL PT COMPARED TO THE PROPOSED APPROACH. THE APPROACHES ARE IMPLEMENTED BASED ON THE THREE TRAINING SETS AND EVALUATED BASED ON THE THREE TEST SETS.

		Test 2014				**Test 2015**				**Test 2016**			
		P@100m	**P@1km**	**P@10km**	**m. error**	**P@100m**	**P@1km**	**P@10km**	**m. error**	**P@100m**	**P@1km**	**P@10km**	**m. error**
Training 2014	SocialSensor [22]	5.87	23.01	39.92	230	-	-	-	-	-	-	-	-
	USEMP [41]	1.61	23.50	40.80	168	-	-	-	-	-	-	-	-
	UQ-DKE [4]	4.98	19.56	41.71	51	-	-	-	-	-	-	-	-
	TALP-UPC [13]	4.12	16.53	34.33	84	-	-	-	-	-	-	-	-
	RECOD [31]	6.06	21.03	37.59	233	-	-	-	-	-	-	-	-
	ICSI/TU Delft [7]	3.15	16.65	34.70	307	-	-	-	-	-	-	-	-
	proposed approach	6.30	25.20	45.64	26	-	-	-	-	-	-	-	-
Training 2015	Baseline [53]	4.26	18.74	40.43	61	4.23	18.44	39.96	71	-	-	-	-
	CERTH/CEA LIST [26]	6.43	24.71	43.57	60	6.40	24.33	43.07	69	-	-	-	-
	ImCube [19]	1.84	8.68	21.34	280	1.84	8.56	21.07	293	-	-	-	-
	Geo_ML [12]	1.26	9.73	30.80	271	1.25	9.51	30.03	291	-	-	-	-
	RECOD [30]	5.61	20.01	37.03	292	5.49	19.75	36.60	310	-	-	-	-
	proposed approach	6.39	**25.45**	45.68	27	6.39	25.13	45.26	30	-	-	-	-
Training 2016	Baseline [53]	3.90	18.31	39.72	70	3.85	17.99	39.28	81	3.82	17.73	39.06	80
	CERTH/CEA LIST [27]	**6.51**	25.01	43.80	55	**6.57**	24.84	43.36	70	**6.43**	24.55	43.32	65
	RECOD [31]	6.29	21.58	38.73	227	6.54	22.06	38.89	234	6.06	21.01	37.91	259
	UoA [45]	2.91	14.97	35.41	87	2.91	14.70	35.58	99	2.88	14.12	35.24	94
	proposed approach	6.23	25.26	**45.76**	**25**	6.33	**25.16**	**45.33**	**29**	6.22	**24.85**	**45.39**	**28**
YFCC	**proposed approach**	7.60	27.91	48.44	14	7.66	27.79	47.87	16	7.52	27.40	47.86	16

this work. However, as expected, there were also numerous cases, where the approach failed. We have found that this was mainly due to a large mismatch between the Flickr-based LM and the language used on Twitter, as well as to the highly irregular and dynamic linguistic patterns arising on Twitter. Due to space limitations, we cannot provide further details on these tests in this paper, and we leave as future work, a comprehensive performance evaluation of geotagging across datasets from different sources, as well as on new methods to improve cross-dataset performance.

V. Conclusions

We presented a text-based geotagging approach that improves upon previous Language Model-based systems thanks to a number of novel feature selection and weighting schemes. The proposed approach was shown to consistently outperform or be highly competitive to the state-of-the-art through comprehensive experiments on four editions of the Mediaeval Placing Task, a popular open benchmark for the multimedia community, while several performance aspects of the proposed approach were examined through carefully designed experiments. By releasing the source code of the proposed approach along with the best performing Language Models that we generated, we aspire to provide a very strong and robust state-of-the-art method to be used for comparisons, and to stimulate further research on the problem.

Acknowledgements. This work is supported by the REVEAL and InVID projects, partially funded by the European Commission under contract numbers 610928 and 687786.

References

[1] E. Amitay, N. Har'El, R. Sivan, and A. Soffer, "Web-a-where: geotagging web content," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2004, pp. 273–280.

[2] J. Cao, "Photo Set Refinement and Tag Segmentation in Georeferencing Flickr Photos." in *MediaEval*, 2013.

[3] J. Cao, Z. Huang, and Y. Yang, "Spatial-aware multimodal location estimation for social images," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 119–128.

[4] J. Cao, Z. Huang, Y. Yang, and H. T. Shen, "UQ-DKE's Participation at MediaEval 2014 Placing Task." in *MediaEval*, 2014.

[5] J. Choi, C. Hauff, O. Van Laere, and B. Thomee, "The placing task at mediaeval 2015," in *MediaEval 2015, Wurzen, Germany, 14-15 September 2015*. CEUR, 2015.

[6] ——, "The placing task at mediaeval 2016," in *MediaEval 2016, Hilversum, The Netherlands, October 20-21, 2016*. CEUR, 2016.

[7] J. Choi and X. Li, "The 2014 ICSI/TU Delft Location Estimation System." in *MediaEval*, 2014.

[8] J. Choi, B. Thomee, G. Friedland, L. Cao, K. Ni, D. Borth, B. Elizalde, L. Gottlieb, C. Carrano, R. Pearce *et al.*, "The placing task: A large-scale geo-estimation challenge for social-media videos and images," in *Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia*. ACM, 2014, pp. 27–31.

[9] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 761–770.

[10] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes, "Improving efficiency and accuracy in multilingual entity extraction," in *I-SEMANTICS 2013 - 9th International Conference on Semantic Systems, ISEM '13, Graz, Austria, September 4-6, 2013*, 2013, pp. 121–124.

[11] J. Davies, J. Hare, S. Samangooei, J. Preston, N. Jain, D. Dupplaw, and P. H. Lewis, "Identifying the geographic location of an image with a multimodal probability density function," in *MediaEval*, 2013.

[12] N. Duong-Trung, M. Wistuba, L. R. Drumond, and L. Schmidt-Thieme, "Geo_ML @ MediaEval Placing Task 2015," in *MediaEval*, 2015.

[13] D. Ferrés and H. Rodríguez, "TALP-UPC at MediaEval 2014 Placing Task: Combining Geographical Knowledge Bases and Language Models for Large-Scale Textual Georeferencing." in *MediaEval*, 2014.

[14] C. Hauff and G.-J. Houben, "Geo-location estimation of flickr images: social web based enrichment," in *European Conference on Information Retrieval*. Springer, 2012, pp. 85–96.

[15] C. Hauff, B. Thomee, and M. Trevisiol, "Working Notes for the Placing Task at MediaEval 2013." in *MediaEval*, 2013.

[16] J. Hays and A. A. Efros, "IM2GPS: estimating geographic information from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008*. IEEE, 2008, pp. 1–8.

[17] ——, "Large-scale image geolocalization," in *Multimodal Location Estimation of Videos and Images*. Springer, 2015, pp. 41–62.

[18] C. F. Karney, "Algorithms for geodesics," *Journal of Geodesy*, vol. 87, no. 1, pp. 43–55, 2013.

[19] P. Kelm, S. Schmiedeke, and L. Goldmann, "Imcube@ MediaEval 2015 Placing Task: A Hierarchical Approach for Geo-referencing Large-Scale Datasets," in *MediaEval*, 2015.

[20] P. Kelm, S. Schmiedeke, and T. Sikora, "A hierarchical, multi-modal approach for placing videos on the map using millions of flickr pho-

tographs," in *Proceedings of the 2011 ACM workshop on Social and behavioural networked media access*. ACM, 2011, pp. 15–20.

[21] C. Keßler, K. Janowicz, and M. Bishr, "An agenda for the next generation gazetteer: Geographic information contribution and retrieval," in *Proceedings of the 17th ACM SIGSPATIAL int. conference on advances in Geographic Information Systems*. ACM, 2009, pp. 91–100.

[22] G. Kordopatis-Zilos, G. Orfanidis, S. Papadopoulos, and Y. Kompatsiaris, "SocialSensor at MediaEval Placing Task 2014." in *MediaEval*, 2014.

[23] G. Kordopatis-Zilos, S. Papadopoulos, and Y. Kompatsiaris, "Geotagging social media content with a refined language modelling approach," in *Pacific-Asia Workshop on Intelligence and Security Informatics*. Springer, 2015, pp. 21–40.

[24] ——, "In-depth exploration of geotagging performance using sampling strategies on yfcc100m," in *Proceedings of the 2016 ACM Workshop on Multimedia COMMONS*. ACM, 2016, pp. 3–10.

[25] G. Kordopatis-Zilos, S. Papadopoulos, E. S. Xioufis, A. L. Symeonidis, and Y. Kompatsiaris, "CERTH at MediaEval Placing Task 2013." in *MediaEval*, 2013.

[26] G. Kordopatis-Zilos, A. Popescu, S. Papadopoulos, and Y. Kompatsiaris, "CERTH/CEA LIST at MediaEval Placing Task 2015," in *MediaEval*, 2015.

[27] ——, "Placing Images with Refined Language Models and Similarity Search with PCA-reduced VGG Features," in *MediaEval*, 2016.

[28] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer, "DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015. [Online]. Available: http://dx.doi.org/10.3233/SW-140134

[29] L. T. Li, J. Almeida, O. A. B. Penatti, R. T. Calumby, D. C. G. Pedronette, M. A. Gonçalves, and R. da Silva Torres, "Multimodal Image Geocoding: The 2013 RECOD's Approach." in *MediaEval*, 2013.

[30] L. T. Li, J. A. Muñoz, J. Almeida, R. T. Calumby, O. A. Penatti, Í. C. Dourado, K. Nogueira, P. R. M. Júnior, L. A. Pereira, D. C. Pedronette *et al.*, "RECOD @ Placing Task of MediaEval 2015," in *MediaEval*, 2015.

[31] L. T. Li, O. A. B. Penatti, J. Almeida, G. Chiachia, R. T. Calumby, P. R. Mendes-Junior, D. C. G. Pedronette, and R. da Silva Torres, "Multimedia Geocoding: The RECOD 2014 Approach." in *MediaEval*, 2014.

[32] X. Li, M. A. Larson, and A. Hanjalic, "Geo-distinctive visual element matching for location estimation of images," *arXiv preprint arXiv:1601.07884*, 2016.

[33] X. Li, M. Riegler, M. Larson, and A. Hanjalic, "Exploration of Feature Combination in Geo-visual Ranking for Visual Content-based Location Prediction," in *MediaEval*, 2013.

[34] M. D. Lieberman, H. Samet, and J. Sankaranayananan, "Geotagging: using proximity, sibling, and prominence clues to understand comma groups," in *proceedings of the 6th workshop on geographic information retrieval*. ACM, 2010, p. 6.

[35] T.-Y. Lin, S. Belongie, and J. Hays, "Cross-view image geolocalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 891–898.

[36] J. Luo, D. Joshi, J. Yu, and A. Gallagher, "Geotagging in multimedia and computer visiona survey," *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 187–211, 2011.

[37] S. E. Middleton and V. Krivcovs, "Geoparsing and geosemantics for social media: spatio-temporal grounding of content propagating rumours to support trust and veracity analysis during breaking news," *ACM Transactions on Information Systems*, vol. 34, no. 3, pp. 1–27, 2016.

[38] J. A. Muñoz, L. T. Li, Í. C. Dourado, K. Nogueira, S. G. Fadel, O. A. Penatti, J. Almeida, L. A. Pereira, R. T. Calumby, J. A. dos Santos *et al.*, "Recod@ placing task of mediaeval 2016: A ranking fusion approach for geographic location prediction of multimedia objects," in *MediaEval*, 2016.

[39] N. O'Hare and V. Murdock, "Modeling locations with social media," *Information Retrieval*, vol. 16, no. 1, pp. 30–62, 2013.

[40] A. Popescu, "CEA LIST's Participation at MediaEval 2013 Placing Task." in *MediaEval*, 2013.

[41] A. Popescu, S. Papadopoulos, and I. Kompatsiaris, "USEMP at MediaEval Placing Task 2014." in *MediaEval*, 2014.

[42] B. D. Ripley, *Spatial statistics*. John Wiley & Sons, 1981.

[43] P. Serdyukov, V. Murdock, and 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*. ACM, 2009, pp. 484–491.

[44] B. W. Silverman, *Density estimation for statistics and data analysis*. CRC press, 1986, vol. 26.

[45] S. K. Singh and D. Rafiei, "Geotagging flickr photos and videos using language models," in *MediaEval*, 2016.

[46] P. D. Smart, C. B. Jones, and F. A. Twaroch, "Multi-source toponym data integration and mediation for a meta-gazetteer service," in *International Conference on Geographic Information Science*. Springer, 2010, pp. 234–248.

[47] S. Subramanian, V. Vidyasagaran, and K. Chandramouli, "VIT@ MediaEval 2013 Placing Task: Location Specific Tag Weighting for Language Model Based Placing of Images." in *MediaEval*, 2013.

[48] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "YFCC100M: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[49] M. Trevisiol, H. Jégou, J. Delhumeau, and G. Gravier, "Retrieving geolocation of videos with a divide & conquer hierarchical multimodal approach," in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. ACM, 2013, pp. 1–8.

[50] O. Van Laere, J. Quinn, S. Schockaert, and B. Dhoedt, "Spatially aware term selection for geotagging," *IEEE transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 221–234, 2014.

[51] O. Van Laere, S. Schockaert, and B. Dhoedt, "Finding locations of flickr resources using language models and similarity search," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*. ACM, 2011, p. 48.

[52] ——, "Georeferencing flickr photos using language models at different levels of granularity: An evidence based approach," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 16, pp. 17–31, 2012.

[53] ——, "Georeferencing Flickr resources based on textual meta-data," *Information Sciences*, vol. 238, pp. 52–74, 2013.

[54] T. Weyand, I. Kostrikov, and J. Philbin, "Planet-photo geolocation with convolutional neural networks," *arXiv preprint arXiv:1602.05314*, 2016.

[55] W. Zhang and J. Gelernter, "Geocoding location expressions in twitter messages: A preference learning method," *Journal of Spatial Information Science*, vol. 2014, no. 9, pp. 37–70, 2014.

[56] Y.-T. Zheng, Z.-J. Zha, and T.-S. Chua, "Research and applications on georeferenced multimedia: a survey," *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 77–98, 2011.

Giorgos Kordopatis-Zilos received the Diploma degree in Electrical and Computer Engineering in the Aristotle University of Thessaloniki (AUTH), Greece in 2013. Since September 2014, he has been working as a research assistant at the Information Technologies Institute (ITI) of the Centre for Research and Technology Hellas (CERTH). He is currently pursuing his Ph.D degree in distance at Queen Mary University of London, on the topic of near-duplicate video retrieval. His research interests include multimedia analysis, indexing and retrieval, web scale data mining, geographic information retrieval and deep learning. He is a Graduate Student Member of IEEE.

Symeon Papadopoulos received the Diploma degree in Electrical and Computer Engineering in the Aristotle University of Thessaloniki (AUTH), Greece in 2004. In 2006, he received the Professional Doctorate in Engineering (P.D.Eng.) from the Technical University of Eindhoven, the Netherlands. Since September 2006, he has been working as a research associate with the Information Technologies Institute (ITI), part of the Centre for Research and Technology Hellas (CERTH), on a wide range of research areas such as information search and retrieval, social network analysis, data mining and web multimedia knowledge discovery. In 2009, he completed a distance-learning MBA degree in the Blekinge Institute of Technology, Sweden. In 2012, he defended his Ph.D. thesis in the Informatics department of AUTH on the topic of large-scale knowledge discovery from social multimedia. He is currently Chair of the IEEE Special Technical Community on Social Networking (STCSN).

Ioannis (Yiannis) Kompatsiaris is a Senior Researcher (Researcher A) with the Information Technologies Institute / Centre for Research and Technology Hellas, Thessaloniki, Greece. His research interests include semantic multimedia analysis, indexing and retrieval, social media and big data analysis, knowledge structures, reasoning and personalization for multimedia applications, eHealth, security and environmental applications. He received his Ph.D. degree in 3-D model based image sequence coding from the Aristotle University of Thessaloniki in 2001. He is the co-author of 90 papers in refereed journals, 38 book chapters, 8 patents and more than 320 papers in international conferences. He has been the coorganizer of various international conferences and workshops and has served as a regular reviewer for a number of journals and conferences. He is a Senior Member of IEEE and member of ACM.