Retrieving Geo-Location of Videos with a Divide & Conquer Hierarchical Multimodal Approach

Michele Trevisiol [1*] trevi@yahoo-inc.com
Hervé Jégou [2] herve.jegou@inria.fr
Jonathan Delhumeau [2] jonathan.delhumeau@inria.fr
Guillaume Gravier [3] guillaume.gravier@irisa.fr

[1]Web Research Group, Universitat Pompeu Fabra, Barcelona, Spain
[1]Yahoo! Research, Barcelona, Spain
[2]INRIA, Rennes, France
[3]CNRS/IRISA, Rennes, France

ABSTRACT

This paper presents a strategy to identify the geographic location of videos. First, it relies on a multi-modal cascade pipeline that exploits the available sources of information, namely the user's upload history, his social network and a visual-based matching technique. Second, we present a novel divide & conquer strategy to better exploit the tags associated with the input video. It pre-selects one or several geographic area of interest of higher expected relevance and performs a deeper analysis inside the selected area(s) to return the coordinates most likely to be related to the input tags. The experiments were conducted as part of the MediaEval 2012 Placing Task. Our approach, which differs significantly from the other submitted techniques, achieves the best results on this benchmark when considering the same amount of external information, *i.e.* when not using any gazetteers nor any other kind of external information.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Geotags, Location, Placing Task, Video Annotation, Flickr

1. INTRODUCTION

Geotagging is the process of automatically adding geographical identification metadata to media objects, in particular to images and videos. This geo-information is called *geotag(s)*, and usually consists of the latitude and longitude world-map coordinates. Determining the place where the content has been captured dramatically extends the knowledge around the media object, especially when combined with time information. Linking time- and geographical-related content offers a new and practical way of automatically searching, organizing or visualizing personal and professional media databases. It also enables the retrieval of various relevant content such as web pages, hence providing users with a wide variety of location-specific information.

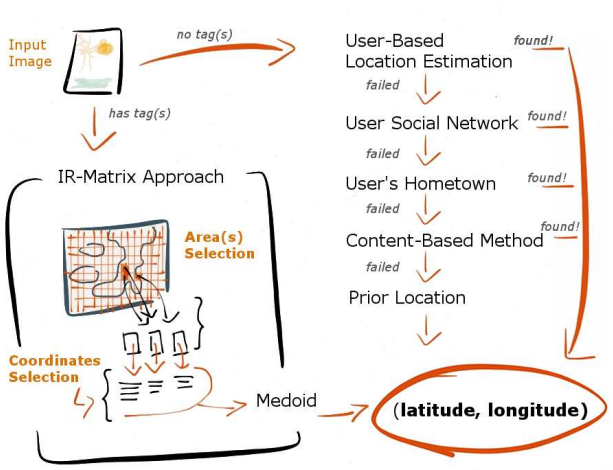


Figure 1: Approach Model Sketch.

In recent years, we have witnessed a dramatic increase in the number of such geotagged media data. Due to the massive spreading of GPS-enabled cameras and mobile phones, geographic coordinates are captured and attached to the content generated by these devices. However, most of the media available remain deprived of this information. For this reason, the problem of automatically assigning geotags to media content (and conversely) is a challenge that increasingly interests the Multimedia community, as reflected by the success of the Mediaeval benchmark's placing task [16]. This paper presents an efficient and effective geotagging system to address this problem, which is illustrated in Figure 1. A multimodal strategy hierarchically processes the sources of information by decreasing order of expected informativeness: Tags, user's upload history and social graph, user's personal information (home town) and visual content. When the most reliable information is missing, the system solely relies on the most informative amongst the remaining sources, with the prior most likely location as a final backup. We do not assume any prior knowledge about the city or country the video was taken in.

Beyond this strategy, a key contribution is the tag analysis technique introduced to extract clues about the location. Our motivation is to identify the tags that are likely to convey some geographical information and to discard the ones

* Work done while visiting PhD student at INRIA Rennes

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ICMR'13, April 16–20, 2013, Dallas, Texas, USA.
Copyright 2013 ACM 978-1-4503-2033-7/13/04 ...$15.00.

that are deemed irrelevant. Indeed, by classifying 54% of the tags from Flickr images using WordNet[1], Sigurbjörnsson and van Zwol [19] observed that 28% of these tags were classified as locations, which suggests that tags have various degree of interest with respect to geotagging.

After filtering out noisy tags in a pre-processing step, we propose two different tag processing techniques, both scalable and not limited to tags that are location names. The first technique implements a text matching rule and serves as a (strong) baseline. The second approach is a radically different tag analysis technique. Based on a divide & conquer strategy, the relationship between tags and coordinates is analyzed to better reflect the informativeness of tags with respect to location.

Our multimodal framework and this new joint tag/location method are the main contributions of this paper. As secondary contributions, we show the interest of different strategies borrowed from other contexts. In particular, we show that the Okapi BM25 weighting scheme [20] is of interest in our context of video localization. Similarly, we have imported some techniques first proposed for image retrieval.

Experiments are carried out on the MediaEval 2012 placing task. Our method is compared with the best participating teams using the official evaluation protocol, and is shown to outperform the state of the art in the same setup, *i.e.*, when using the same official input provided by the organizers. The specific interest of our tag-analysis technique is demonstrated by comparing it to a strong baseline, which by itself already achieves state-of-the-art performance.

The paper is organized as follows. Section 2 describes the background on geotagging, including related work and solutions adopted by Mediaeval participants that we compare with. An overview of our multimodal processing pipeline and its components are presented in Section 3. The new joint tag/location method which is detailed separately in Section 4. The experimental setup and results are shown in Section 5. Section 6 concludes the paper.

2. BACKGROUND

This section makes a brief overview of key approaches and trends for geotagging of images and videos. As this topic has received growing attention in the multimedia, computer vision and social networks communities, we refer to the recent survey by Luo *et al.* [11] for a wider overview of the techniques. We also present the placing task of the MediaEval 2012 benchmark and detail some approaches designed by some participants. They will serve as a comparison benchmark in the experiments.

2.1 Related work

The location of a video or image is typically extracted from two main sources of information, namely the textual information (*i.e.*, tags, title, description) and the visual characteristics (*i.e.*, global/local features). This specific information might be combined with some prior statistics on possible locations. In some situations, the media data is not associated with any textual information, therefore determining the location relies on the only available information, *i.e.*, the image or video itself. In contrast to the techniques exploiting metadata such as tags or EXIF information, these approaches are usually referred to as content-based approaches. In this line of research, Hays and Efros [3] proposed a purely visual approach[2] that estimates the image location as a probability distribution over the Earth's surface. Penatti *et al.* [14] proposed an approach called *bag-of-scenes*. First, they create a dictionary of scenes from places of interest, where each of them can be represented by a certain type of low-level features. Then, each video frame is compared with the dictionary and the most similar scene is selected. This allows associating a scene with each frame in order to save more semantic information.

However, the content-based approach alone is not reliable enough to be considered effective. Multimodal solutions exploit more sources of information, in particular textual information. The words extracted from the available text (*i.e.*, title, description and tags) are often associated to geographic coordinates to determine the most common words for each possible location. An example of such an approach is the work of Serdyukov *et al.* [17], which constructs a $m \times n$ grid based on latitude and longitude, where each cell represents a location. Images whose locations are known are associated with their corresponding grid cells. Finally, a language model is estimated from the tags associated with a particular location, taking into account neighbors' influence and leveraging spatial ambiguity. Sergieh *et al.* [18] worked on the reciprocal problem, proposing a statistical model for automatic image annotation. Given an image with coordinates, they infer some relevant tags based on textual information from images in the database that are physically located nearby and which have similar visual content. Crandall *et al.* [2] used both image content and textual metadata to predict the location of an image at two levels of granularity: city level (about 100km), and at the individual landmark level (about 100m). However they limited their experiments to a specific set of landmarks in a fixed set of cities. In contrast, the problem considered in this paper makes no assumptions on the data set and on the level of granularity in the detection step. O'Hare and Murdock [13] proposed a statistical language modeling approach, also dividing the Earth into grid cells. Their approach is based on the Word-Document paradigm, and they investigate several ways to estimate the models, based on the term frequency and the user frequency.

2.2 MediaEval 2012

MediaEval is an international evaluation campaign in which the Placing Task [16] is dedicated to the geo-localization problem addressed in this paper. The goal is to determine as accurately as possible the location, in terms of latitude and longitude, of a set of Flickr videos. The task covers several cases, called *runs*, each of them being restricted by some constraints on the type of information used. To ensure that the training data is the same for all techniques so as to provide a fair comparison, we focus on techniques that only used the information provided by MediaEval. Hence, extra resources, such as gazetteers (*e.g.*, GeoNames, WordNet) or any kind of external information (*e.g.*, Wikipedia, Google Maps), are excluded from all the experiments so as to focus on the data processing techniques proposed.

2.2.1 *Dataset description*

[1] http://wordnet.princeton.edu/

[2] Note that the dataset they consider only includes images associated with a geotag such as a country, a city or as a touristic site (*e.g.*, "Pisa", "Nikko", "Orlando").

	no tags	single tag	size
Train Set	454,338 (14.2%)	27,488 (0.9%)	3,200,757
Test Set	1,902 (45.5%)	139 (3.3%)	4,182

Table 1: Number of media objects without tags.

The MediaEval 2012 Placing task dataset gathers content from Flickr in Creative Common license and is divided into a train set with both images ($\approx 3.2M$) and videos ($\approx 15K$) and a test set with 4,182 videos, from more than $71K$ users. Metadata is associated with each media object and consists of various information such as ownership (Flickr user id and nickname), timestamps (upload and shot time), textual data (tags/keywords, title and description), social network (owner's contact user ids), comments and favorites (contents and users that made them) and, of course, the latitude and longitude within a certain level of accuracy. Note that in Flickr there are 16 levels of accuracy, from the most general (*i.e.*, country name) to the most specific (*i.e.*, street address). Table 1 summarizes the number of objects with tag(s) associated for each dataset. Clearly, the test set includes a large proportion of videos with no tags. Moreover, as tags are arbitrarily added by users without any constraint or rule, a large proportion of the tags is meaningless. Overall, many annotated objects are not associated with a single useful tag. This makes this benchmark both challenging and realistic.

2.2.2 Evaluation protocol

Our evaluation strictly follows the rules of the MediaEval 2012 placing task. The accuracy of the estimated location is measured by *great circle* distances between the predicted and the actual geo-coordinates encoded in the video. The Haversine distance is used to measure the discrepancy between the estimated location and the real one. The ground-truth is supplied by Flickr users at upload time.

2.2.3 Description of submitted geotagging techniques

Various approaches were taken by MediaEval's participants to address the problem. This section presents some representative methods, including the most successful ones, which are included in the comparison of Section 5. **Choi *et al.* [1] gave priority to the textual information,** using title and tags/keywords, but discarding the description. They computed a geographic spread for each word (in tags and title), similar to what we do. In addition, they exploited the GeoNames database to have a toponym resolution in order to filter out irrelevant words. They also included part-of-speech retrieved to perform more precise filtering using Augmented-WordNet[3]. In case of no candidate coordinates, they used the user's home location, or as last resort, the prior location (*i.e.*, fixed location computed *a priori*). **Li *et al.* [10] extended the successful bag-of-scenes technique** [14], including the histogram of motion patterns. They aggregated with a fusion module both a textual (based on tags, title and description) and a visual approach. Interesting results are presented for the content-based (visual) task, but they are not the main focus of this paper. **Popescu and Ballas [15]** tackled the problem by splitting the Earth surface in small cells of size 0.01 of latitude and longitude degree, characterized by a set of tags and their probability of occurrence in that cell. They selected only pairs of tags with a high probability of occurrence within a smaller radius in order to extract a set of unambiguous pairs of potential toponyms. Then they matched the tags for each test video with the cells of the unambiguous pair (if it is found), or with the whole set of cells, considering as top ranked the selected cells and their neighbors. **Van Laere *et al.* [9]** applied a divide & conquer approach splitting the problem in two phases. Given the test video, in the first step they find the most likely cluster to contain the location with a Naive Bayes classifier. Then with a similarity search, they find the training items whose tags are the closest to the ones of the test video. If the test video has no tags, they use user's hometown, title and description as if they were regular tags. As a last resort, they also used a prior static location. **Kelm *et al.* [8] presented a hierarchical framework** that combines textual and visual features for different granularity. First they divide up the Earth in regions using meridians and parallels, then they generated textual and visual prototypes for each of them. For the textual part, they translated in English tags, title and description, then they extracted words using a NLP approach, and finally they applied a stemmer and a stop-word elimination. Given the test video, they select the region and the images/videos with highest probability to contain the extracted words (using a bag-of-words approach). Then, given a list of ranked candidates, with a visual search they select the most similar.

[3] http://ai.stanford.edu/~rion/swn

3. MULTI-MODAL CASCADE

This section describes our multimodal and hierarchical processing pipeline. It starts with a tag comparison technique based on frequency matching, followed by a description of how the remaining sources of information are processed. As shown in Table 1, many videos in the test set are not described by tags. To handle these cases, we exploit additional information in a pipeline: If one source of information is absent or fail to provide a reliable prediction, the next is considered. The pipeline operates in the following order which was chosen according to the amount of information conveyed by each source, as discussed later in the experimental section: *a*) tags *b*) user's upload history, social information, *c*) user's home town, *d*) content-based matching, *e*) prior-location. This process is illustrated in Fig. 1.

3.1 Tag processing: IR-frequency

The frequency tag processing technique proposed hereafter is the first way we propose to exploit the tags. This technique, which is referred to as *IR-frequency* in the following, mainly serves as a baseline. A better novel technique will be presented in the dedicated Section 4.

3.1.1 Pre-processing

Flickr normalizes the set of raw tags by lower-casing them, removing white space and stop-words, and replacing commas with white space. For example, the set of tags "Trip 2010, Sagrada Familia, Barcelona" becomes "trip2010 sagradafamilia barcelona". Remember that tags are arbitrarily chosen by a user to describe the image. Hence, they might be inconsistent with the image content or location. We further normalized tags so as to defined a set of clean tags, Tc_{train} derived from T_{train}, the entire set of tags in the training data. We removed the accents, discarded numeric tags (almost never relevant for the location), and removed numeric

characters from the alphanumeric tags. A stop-list containing common words (*e.g.*, travel, birthday, cat, geotag, camera) and product or brand names (*e.g.*, iPhone, Canon) was used to filter out non informative tags. So called machine tags[4] (or *mtags*), *i.e.*, one or more tags that Flickr recognized as a location (usually a country name or sometimes a city name), are kept unchanged and will be processed independently of the other tags as they are highly accurate and relevant. Note that after pre-processing, only 39.9 % of the videos contain tags.

3.1.2 Geo-relevance filtering

For a baseline method based on direct tag matching, selecting tags relevant to the geo-location is a crucial step. Apart from machine tags which are deemed relevant, we implemented a geo-relevance filtering based on the geographic spread of a tag in the training data. Figure 2 illustrate this idea by showing how some tags are spread across the globe: Tags specific to a location (bottom row) are mostly concentrated in a small area while others (top row) exhibit a high dispersion.

To select relevant tags in Tc_{train}, we compute for each tag t_i its frequency of occurence f_{t_i} in the training data and the average Haversine distance d_{t_i} between the coordinates of the data which contain t_i. Tags that do not match the following condition

$$\forall t_i \in T_{\text{train}}, \quad t_i \in Tc_{\text{train}} \iff \begin{cases} f_{t_i} \leq 50, \\ d_{t_i} \geq 200. \end{cases}$$

are removed from Tc_{train} where the thresholds were experimentally defined.

3.1.3 Frequency matching

Given the set of tags retained, one can group coordinates associated to the same set of tags. The idea is first, to preselect some set of tags that have at least one mtag in common (if available otherwise a normal tag), and finally to rank each of them by the occurrences of the common (m)tags.

We consider each training document, image or video, as a geo-annotated document described by a set of tags. For each set of tags, including machine tags, we collect all the coordinates from documents described by the same set of tags, along with the number of such documents. For example, for the set of tags "france", "pompidou" and "paris", we collect the following coordinates (48.8611, 2.3521):12, (48.6172, 2.213):3.

Given a test video, if it contains mtags we retrieve all the documents where there is at least one common mtag, otherwise we do the same with tags. Those documents are further ranked according to the number of tags they share with the test video. The top ranked document (or documents in case of equality) is selected and the medoid of all the locations attached to the corresponding set of tags, weighted by the number of occurences of the coordinates, is taken as the test material's geo-coordinate.

3.2 User data processing

When no tags are left after filtering or if no documents in the training data is found with at least one tag in common with the test video, we rely on the user data provided to predict a location.

[4] http://www.flickr.com/groups/api/discuss/72157594497877875/

3.2.1 User upload history

For each user with images or videos in the training set, we picked a pre-computed *user location* based on the most frequent location for his content. We found that 35.6 % of the users in the test set appear in the training data. Assuming that users tend to visit the same places more than once, we seek to exploit the documents previously uploaded. For each user in the training set, we compute the medoid of the geo-coordinates of all its training data. The obtained location is used as geo-coordinate when tag-based geotagging fails. We observed that using the user prior location significantly improved the results.

3.2.2 Social network extension

For users not present in the training data, we make use of their social connections to infer a potential location. The idea is to find the user locations of all the contacts and use the medoid as the most likely location for the test video. This general idea is refined based on the groups which are used in Flickr user connections, namely *family*, *friends*, and *contacts*. We assume that *family* is closer to *friends* which in turn is closer to *contacts* and process the groups in that specific order. If the user has enough connections in one group, then the video location is obtained from the contacts in the group. Else, we move on to the next group.

Using both user upload history and social network extension, 79% of the test videos are covered.

3.2.3 User hometown

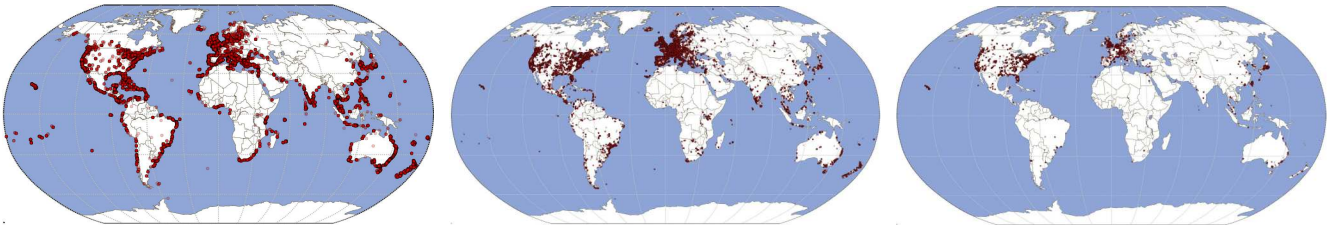
In case neither upload history nor social connections are available, the hometown of the user, as given by its Flickr profile is used. When available, the hometown is given as a place name, *e.g.*, "San Francisco, California, United States", rather than as coordinates. We process the hometown information as if they were tags describing the test video, applying the same process as described in Section 3.1.3 to determine geo-coordinates. Note however that the user hometown is not always well specified (*i.e.*, only the state or the country is specified) and is not always precise (*e.g.*, with very large cities like New York, the estimated coordinates can be very far from the real ones).

3.3 Content-based processing

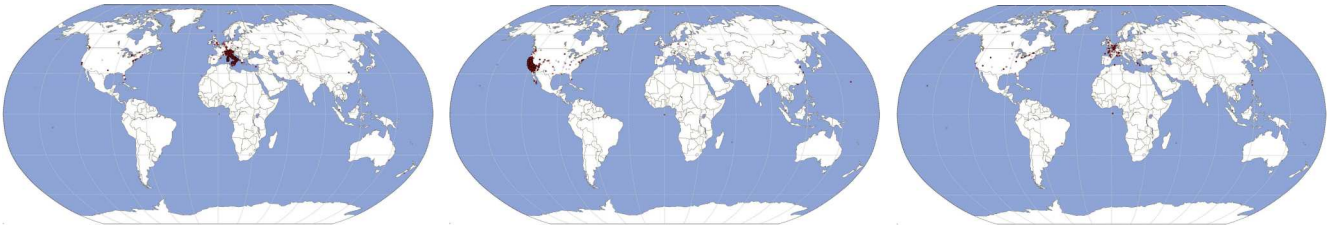
Content-based geo-tagging exploiting image matching is finally used. However, the input video set is not large enough with respect to the total number of locations, and include many indoor scenes. Therefore the visual approach, which requires the same views of a given location, is less important than other sources of information. Anyway, for this purpose each keyframe or image is described based on SIFT local descriptors computed over a dense grid. A power law of 0.5 is applied before L2 normalization [7]. PCA and whitening are applied before aggregating vectors into a global high-dimensionality VLAD descriptor [7] which is reduced to dimension 1,024 by PCA, whitened and normalized. An index is built from those descriptors using product quantization [6] which enables fast approximate nearest neighbor search on all of the test keyframes. For each query, we get the coordinates of the best candidate keyframes and return their medoid.

3.4 When all elses fail...

As a last chance, if all elses fail, we assign a default prior



(a) *beach* (b) *nature* (c) *iphone*



(d) *italy* (e) *california* (f) *paris*

Figure 2: Coordinates of six tags plotted on the world map. The first row shows the spreading of three tags that are not locations. The second one shows respectively a country, a state, and a city.

location regardless of the content. We experimented two strategies for the default location: The medoid of all locations in the training data—which falls near Tokyo—or the medoid of all locations from the training data with no tags—which falls in London.

4. HIERARCHICAL METHOD

The tag processing described in the previous section remains limited and we seek to develop a new approach better exploiting the information conveyed by the tags. In the following, we propose a hierarchical approach based on the vector-space model, using the divide & conquer paradigm, to infer the relationship between tags and locations. This method will be referred to as *IR-matrix* in the following, by analogy to the word-document matrix analysis performed in some information retrieval techniques.

The tags of a given test video are considered as a *query vector*. The idea is to first determine the approximate geographic area in which the video is likely to belong, and to find in turn the most probable coordinates from the known locations in that area. The geographic areas are arbitrarily defined by quantifying the coordinates on a grid, where each cell of the grid is described by a vector of tag weights according to the tag relevances to the area considered. Each geographic area is further defined by a set of geo-coordinates also described with a specific tag vector. The test document is represented as a weighted vector of tags from which the most likely areas and the most likely coordinates are determined.

The steps described in this section replaces the tag filtering and frequency matching in the pipeline described in the previous section. The other steps are left unchanged.

4.1 Tag weighting

Describing a geographic area or a specific coordinate as a set of tags with weights require a weighting scheme that reflects the relationship between tags and coordinates. Similar quantities as those used for tag filtering in section 3.1.2, namely frequency and average distance, are used to measure the relevance of a tag. Rather than eliminating non relevant tags as before, a low weight is given. The following heuristics was used to identifying how *geo-descriptive* a tag is: $\forall t_i \in Tc_{\text{train}}$

$$w_{t_i} = \begin{cases} -1 & \text{if } f_{t_i} > 100K \text{ or } d_{t_i} < 0.2 \\ 10 & \text{if } f_{t_i} \geq 200 \text{ and } 10 \leq d_{t_i} \leq 50 \\ 5 & \text{if } f_{t_i} \geq 150 \text{ and } d_{t_i} \leq 70 \\ 1 & \text{otherwise} \end{cases}$$

This weighting was designed to assign higher weights to tags representing geographic information, *i.e.*, not only places but also references to locations such as monuments. Figure 3 shows an example of tags with the highest weight ($w_{t_i} = 10$) sorted by f_{t_i} as opposed to tags sorted by frequency only. All tags with high weights clearly designate locations. Figure 6 shows some examples of weighted query tags.

4.2 Finding the areas

Given a set of tags from a test video, we first want to identify the most likely geographic area(s).

Areas were defined by quantifying the coordinates on a cell grid of 0.1°, *i.e.* a coordinates with latitude 41.12 and longitude −1.23 belongs to the area identified as (41.1,−1.2). Though not the most compact representation, quantization on a cells grid is computationally not expensive. Each cell j in the grid is described by a vector where each bin corresponds to a tag with the corresponding weight defined as:

$$w'_{t_i,j} = f_{t_i,j} w_{t_i,j} \quad (1)$$

where $f_{t_i,j}$ is the number of occurences of t_i in the area j, and $w_{t_i,j}$ is defined from $f_{t_i,j}$ as in section 4.1.

The set of areas is thus represented by a matrix whose rows correspond to tags and whose columns correspond to the geographic areas. The Okapi weighting scheme is applied to all entries in the area matrix—see section 4.4 for details—before smoothing using signed SQRT and L2 normalization, generalizing to text features results from image processing [5]. The area that best fits a test image represented as a vector of tag weights is obtained by multiplying

Tag	Frequency	Average Distance
geotagged	81948	183.600974549
water	36645	394.109273724
beach	35799	189.702712644
nature	35791	330.458447429
2009	35411	158.518797983
2008	34788	163.01576923
california	34756	17.3347855177
2007	33791	163.774148222
sky	32268	500.652309767
travel	29636	227.985879656
usa	27933	112.840988813
italy	26432	28.2242237023
sea	25435	264.814684265
france	25128	30.1802960837
sunset	24839	530.829445257
landscape	24291	399.483239901
snow	24251	252.464879232
europe	24173	121.315134058
blue	23187	554.374093842
2006	22802	188.297981051
australia	22746	48.9602486282
night	22745	439.445587846
germany	22060	30.3288003332
winter	21605	264.227771688
tree	21335	626.460930315
canada	20611	44.1591343262
spain	20583	33.9538296347
bw	20475	504.583508417
architecture	20155	331.577365854
japan	19859	33.8409323297
green	19822	585.402442074
uk	19675	33.246712809
clouds	19597	617.251967179
flower	19563	631.402098385
england	19477	25.9678916173
park	19245	289.55492473
vacation	18870	220.652637728

Tag	Frequency	Average Distance
california	34756	17.3347855177
italy	26432	28.2242237023
france	25128	30.1802960837
australia	22746	48.9602486282
germany	22060	30.3288003332
canada	20611	44.1591343262
spain	20583	33.9538296347
japan	19859	33.8409323297
uk	19675	33.246712809
england	19477	25.9678916173
espana	14740	36.3307740002
scotland	14418	20.0878818974
italia	14266	29.7714336536
deutschland	12469	26.6710348635
mexico	11248	44.1710727199
washington	10528	29.2185701167
texas	9811	26.8342797456
florida	9773	19.2311687716
newyork	9550	26.7850834626
portugal	9005	36.4910584785
switzerland	8842	19.0660337254
sweden	8825	43.4298253824
taiwan	8614	8.58611101877
ireland	8277	27.8336567598
newzealand	7804	34.7207757884
greece	7734	28.8041197511
ontario	7684	13.5533928009
oregon	7501	21.9655577751
unitedkingdom	7469	45.7784858581
netherlands	7258	23.4561091895
austria	6670	17.0680578585
colorado	6659	29.0537398392
thailand	6508	23.5652696046
nsw	6231	28.5942670687
arizona	6041	26.381253062
sanfrancisco	6021	41.1354079352
nyc	5917	32.7179331764

Figure 3: On the left side of the line, there are listed tags before the weighting scheme is applied. On the right side instead, there are shown the tags with highest score ($w_{t_i} = 10$). Both of lists are sorted by term frequency (tf_{t_i}).

the query vector by the area matrix, thus providing a ranked list of areas. The area with the highest matching score is selected, several areas being selected in case of equality.

4.3 Finding the coordinates

Given a selected area, we proceed to find the most likely coordinates for the tags of the test video, following the same principle as before. Similarly to what is done for areas, a tag/coordinate matrix is used to represent coordinates within each area, where each row corresponds to a tag and each column corresponds to a coordinate in the area cell. The weights in the matrix are obtained following the same procedure as for the area matrix, with tag frequencies computed for each coordinate. Okapi weighting, smoothing and L2 normalization are also applied. Given a test query obtained from the tags of the test videos, a ranked list of coordinates is obtained within each of the areas selected in the previous step. The best ranked coordinates are selected from each of the ranked lists and the medoid is used as the geo-coordinates for the test video.

4.4 Tuning Okapi BM25

While tf-idf is commonly employed as a weighting scheme for text representation in the vector-space model, the Okapi BM25 weighting scheme was experimentally found to perform better in our case, confirming previous results [20]. The Okapi weighs are defined as

$$W_{BM}(j,t_i) = \sum IDF(t_i) \times \frac{w'_{t_i,j} \times (k+1)}{w'_{t_i,j} + k \times (1 - b + b \times \frac{|D|}{avg_{dl}})}$$

where $w'_{t_i,j}$ is defined by Eq. 1, avg_{dl} is the average number of tags per training sample, k and b are free parameters usually chosen as $k \in [1.2, 2.0]$ and $b = 0.75$ [12]. The IDF part instead is given by

$$IDF(q_i) = \log \frac{N - N_{t_i} + 0.5}{N_{t_i} + 0.5}$$

where N is the total number of training samples, and N_{t_i} is the number of samples containing tag t_i.

We experimented different values of k, both for area selection and coordinates selection. Contrary to the conclusions of Whissell *et al.* [20], where large values of k ($k \geq 20$) improve the results, we found that small values of k performed better in our case. Table 2 shows some results in terms of coordinates correctly identified for various values of k. While large values decrease performance, small values of k tend to increase the accuracy at a small radius (*i.e.*, 1 km). Various combination of k, for the coarse grain area selection and the fine grain coordinate selection where tested, results being reported in table 3. Combining small values of k improves for both the 1 km and 10 km radii.

radius(km)	**0.001**	**0.01**	**0.1**	**1.2**	**10**	**20**
1	**756**	749	752	720	714	713
10	1626	**1641**	1627	1601	1587	1582
100	2071	2086	**2095**	2085	2071	2068
1000	2737	2739	2751	**2760**	2763	2760
10000	3885	3884	3890	**3892**	3889	3891

Table 2: Estimating values of k_1 for the step of selection of the area, comparing different values of k_1 from 0.001 to 20. For each radius (in km) the correctly detected coordinates for the test videos are counted.

radius(km)	$\frac{k_{1,1}=0.001}{k_{1,2}=0.001}$	$\frac{k_{1,1}=0.001}{k_{1,2}=1.2}$	$\frac{k_{1,1}=0.1}{k_{1,2}=0.1}$	$\frac{k_{1,1}=1.2}{k_{1,2}=1.2}$
1	**786**	756	752	720
10	**1635**	1626	1628	1601
100	2071	2065	**2091**	2079
1000	2759	2753	2769	**2774**
10000	3962	3959	**3964**	**3964**

Table 3: Estimating values of k_1 for first ($k_{1,1}$, selection of the area) and for second step ($k_{1,2}$, selection of coordinates). Where for each radius (in km) the correctly detected coordinates for the test videos are counted.

5. EXPERIMENTS

We evaluate our approaches with the dataset from MediaEval's 2012 Placing Task described in Section 2.2. Our system was trained using the $3.2M$ geotagged images and videos released by the organizers. As shown as an outcome of the 2011's campaign, the tag information is the most reliable one. However, in many situations, a large proportion of the videos have no tag after our filtering steps, for instance about 60% on the Mediaeval benchmark. That is why this section first discusses the respective interests of the other sources of information, which led us to determine the order of priority in our cascade multi-modal approach. We then present how our system performs on the Placing Task of MediaEval 2012, and shows the large improvement brought the IR-Matrix method of Section 4 compared to the baseline tag method and to the submitted techniques.

5.1 Sources of information

Section 3 introduced the secondary sources of information that we exploit when the test video is not associated with any tag after the filtering step. In our cascade architecture, an important choice is the order in which the corresponding

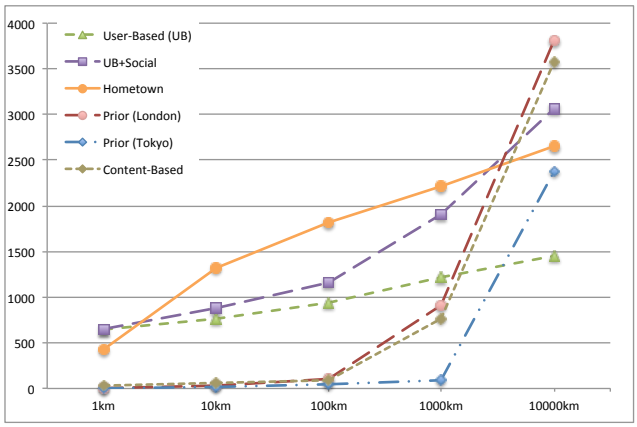


Figure 4: Cumulative values of correctly detected locations for pipeline methods: number of video founds (y-axis) in a radius of x km (x-axis).

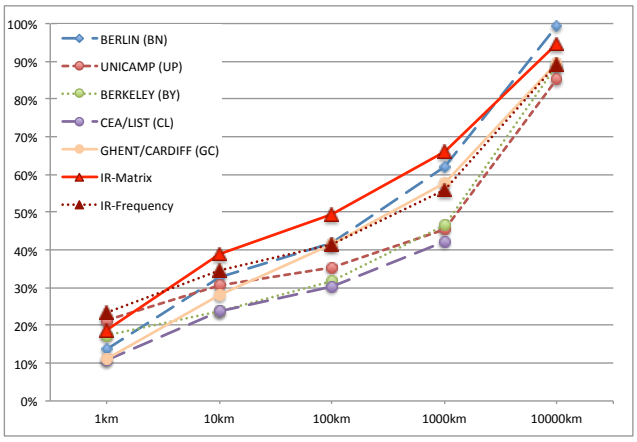


Figure 5: Cumulative correctly detected locations: rate of video founds (y-axis) in a radius of x km (x-axis).

methods appear in the pipeline, as this order impacts the final quality of the system. For this purpose, we have evaluated the respective geo-tagging accuracy provided by each component of our system[5].

Figure 4 shows the number of correctly identified locations for varying radius and for the different sources, except for the visual search, which provides inferior results. For example, in the 1 km radius, the user-based estimation identifies more than 600 video locations, while the estimation with Hometown finds less than 500 of them. However, for radius higher than 1 km, the interest of the Hometown improves and gives the best results among the secondary sources of information. Another observation is the social connections are useful and significantly improve an estimation based only on the other metadata related to the user. The content-based approach is performing poorly due to the types of test videos that contain mainly indoor scenes. However for the smaller radii (1km and 10km) it is slightly better than the prior location, and for this reason in our pipeline approach it is used before. The prior location does not use any information about the query and, as to be expected, leads to a very imprecise estimation which only impacts the 1000 and 10000 km precision measures. Interestingly, London and Tokyo give very different performances. However, in our opinion, this prior information is not really interesting for a real application, as it is not related to a particular video.

The fact that the Hometown gives the best results for radius higher than $1km$ suggests that it should be used as the primary alternative to the tag-based method. However, when combining the different sources in cascade, our preliminary experiments showed that it is worth exploiting user-based and social information first.

To conclude this discussion, our final framework uses the UB+Social as a primary alternative to the tag-based method (See Section 3). If this fails to output a location, we use the Hometown estimation instead if provided, else the visual search engine. The prior location is used as a final backup.

5.2 IR-MATRIX Evaluation

This Section compares the results of our approach to the ones shown by the participants at MediaEval 2012 Placing Task. In addition, we separately show the interest of the IR-Matrix technique introduced in Section 4 over the IR-frequency baseline (See Section 3). These two tag-based approaches are evaluated using the same pipeline, *i.e.*, in the same multi-modal cascade. Therefore, the results of these methods are directly comparable[6].

Figure 5 shows that, overall, our IR-Matrix technique performs the best except in two cases: The last radius, 10000 km, mainly depends on the Prior location and has arguably no practical interest. As mentioned in footnote, the 1 km measurement is not reliable because it is impacted by the artifacts of the train/test duplicates.

Discussion. Among the various methods submitted to Mediaeval'2012, different textual filters have been used. BY [1] computes a geographic spreading based on the spatial variance distribution filtering tags with high variance. BN [8] performs a more complex procedure, translating everything (also title and description) in English, applying a stemming and stop-word filtering, and finally extracting words with a NLP approach. CL [15] works only with tags that have been used by at least two users, and considers only pairs of unambiguous pre-computed toponyms.

Our IR-Matrix technique is less restrictive than these technique because it does not discard any tag, except the ones filtered by a common stop-word list. Instead, it automatically assigns different weights to each of them, which is less radical than the techniques mentioned above and leads to exploit more tags, thereby reducing the information loss. In addition, by considering each cell of the Earth grid as a separate "document", our Word-Document matrix-based approach better identifies the relationship between tags and localization, which in turn provides a useful measure of geo-informativeness to tags.

Concerning the secondary methods, only CL exploited the user's previous uploads in the case where no tags is associated to the test video, and nobody used the social information in order to expand this knowledge. This gives a slight

[5]Note that this evaluation of the respective interest of information sources was first done on the 2011's Mediaeval campaign, without knowing the 2012's test data.

[6]There is a bias for the 1 km radius measure, as some test images were also included in the training set. This basically favors the baseline approach for this measure (and other systems) because one could match some test videos perfectly based on irrelevant tags. We have not exploited this knowledge in our system.



piazzabra **verona italy** veneto northernitaly worldheritagesite unesco unescoworldheritagesite fountain waterfeature ~~video~~ videoclip



provence southoffrance **france** ~~bouchesdurhoneandaimes~~ bouchesdurhone **provencealpescotedazur** aixenprovence coursmirabeau fountain ~~video~~ videoclip ~~lafontaincdeauchaude~~ ~~thehotwaterfountain~~ ~~mossyfountain~~ hotspring ~~moussue~~ ~~bagniers~~ ~~aixcentre~~ aix



norfolk england unitedkingdom greatbritain westnorfolk hunstanton beach sea thewash northsea ~~northpromenade~~ cliff cliffs ~~video~~ videoclip chalk redchalk carstone whitechalk ~~hunstantonformation~~ ~~newhunstanton~~ ~~hunstantonstedmunds~~ **eastanglia**

Figure 6: Query examples with tags. Lightness and size indicate initial weights (w_{t_i} = **10**, 5 or 1), ignored tags are striked-through (~~stop words~~ or ~~not in the database~~).

improvement which is exploited in both our IR-matrix and IR-frequency methods.

6. CONCLUSIONS

This paper introduces a novel system for geo-tagging videos which significantly outperforms techniques of the state of the art, as demonstrated by our experiments performed on the last Mediaeval benchmark.

A key contribution is the novel IR-Matrix location/tag technique based on the Divide & Conquer paradigm, which is simply and efficiently implemented by (query)vector-matrices multiplications. It first provides an estimation of the area of interest, which is then used to determine more precise coordinates that best match the input set of tags. It significantly outperforms a more conventional tag-vector matching technique, such as our IR-Frequency baseline which first detects all the images and videos that contain the specific tag(s), and then selects the one with the highest number of matches. As a complementary technique, we show the interest of the Okapi weighting scheme in this context.

When no reliable tag is available, our processing cascade allows our system to make a prediction based on other sources of information, such as user-related metadata or visual content. To our knowledge, our system is also the first to exploit the social connections for this geo-tagging task.

Although we only considered the meta-data provided in the Mediaeval benchmark, *i.e.*, the Flickr data associated with the videos, we believe that integrating external sources of information, such as a gazetteer, should further improve the overall localization performance, as demonstrated by other works in the field.

7. ACKNOWLEDGMENTS

This work was partially funded by OSEO, French state agency for innovation, in the framework of the Quaero project and by Grant TIN2009-14560-C03-01 of the Ministry of Science and Innovation of Spain. Furthermore, we would like to thank Vincent Claveau for his helpful suggestions.

8. REFERENCES

[1] J. Choi, G. Friedland, V. Ekambaram, and K. Ramchandran. The 2012 ICSI/Berkeley Video Location Estimation System. In *MediaEval*, 2012.

[2] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the World's Photos. In *WWW*, 2009.

[3] J. Hays and A. A. Efros. IM 2 GPS : estimating geographic information from a single image. In *CVPR*, 2008.

[4] M. Jain, R. Benmokhtar, P. Gros, and H. Jégou. Hamming Embedding Similarity-based Image Classification. In *ICMR*, Jun. 2012.

[5] H. Jégou and O. Chum. Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening. In *ECCV*, Oct. 2012.

[6] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *PAMI*, 33(1), Jan. 2011.

[7] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *PAMI*, Sep. 2012.

[8] P. Kelm, S. Schmiedeke, and T. Sikora. How Spatial Segmentation improves the Multimodal. In *MediaEval*, 2012.

[9] O. V. Laere, S. Schockaert, and J. Quinn. Ghent and Cardiff University at the 2012 Placing Task. In *MediaEval*, 2012.

[10] L. Li, J. Almeida, and D. Pedronette. A Multimodal Approach for Video Geocoding. In *MediaEval*, 2012.

[11] J. Luo, D. Joshi, J. Yu, and A. Gallagher. Geotagging in multimedia and computer vision–a survey. *Multimedia Tools Appl.*, 51(1), Jan. 2011.

[12] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[13] N. O'Hare and V. Murdock. Modeling locations with social media. *Information Retrieval*, Apr. 2012.

[14] O. A. B. Penatti, L. T. Li, J. Almeida, and R. da S. Torres. A Visual Approach for Video Geocoding using Bag-of-Scenes. In *ICMR*, 2012.

[15] A. Popescu and N. Ballas. CEA LIST's Participation at MediaEval 2012 Placing Task. In *MediaEval*, 2012.

[16] A. Rae and P. Kelm. Working Notes for the Placing Task at MediaEval 2012. In *MediaEval*, 2012.

[17] P. Serdyukov, V. Murdock, and R. van Zwol. Placing flickr photos on a map. In *SIGIR*, May 2009.

[18] H. M. Sergieh, G. Gianini, M. Döller, H. Kosch, E. Egyed-Zsigmond, and J.-M. Pinon. Geo-based Automatic Image Annotation. In *ICMR*, 2012.

[19] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW*, 2008.

[20] J. Whissell and C. Clarke. Improving document clustering using Okapi BM25 feature weighting. *Information Retrieval*, 14, 2011.