Multimodal Geo-tagging in Social Media Websites using Hierarchical Spatial Segmentation

Pascal Kelm
Communication Systems Group
Technische Universität Berlin
Germany
kelm@nue.tu-berlin.de

Sebastian Schmiedeke
Communication Systems Group
Technische Universität Berlin
Germany
schmiedeke@nue.tu-berlin.de

Thomas Sikora
Communication Systems Group
Technische Universität Berlin
Germany
sikora@nue.tu-berlin.de

ABSTRACT

These days the sharing of photographs and videos is very popular in social networks. Many of these social media websites such as Flickr, Facebook and Youtube allows the user to manually label their uploaded videos with geo-information using a interface for dragging them into the map. However, the manually labelling for a large set of social media is still borring and error-prone. For this reason we present a hierarchical, multi-modal approach for estimating the GPS information. Our approach makes use of external resources like gazetteers to extract toponyms in the metadata and of visual and textual features to identify similar content. First, the national borders detection recognizes the country and its dimension to speed up the estimation and to eliminate geographical ambiguity. Next, we use a database of more than 3.2 million Flickr images to group them together into geographical regions and to build a hierarchical model. A fusion of visual and textual methods for different granularities is used to classify the videos' location into possible regions. The Flickr videos are tagged with the geo-information of the most similar training image within the regions that is previously filtered by the probabilistic model for each test video. In comparison with existing GPS estimation and image retrieval approaches at the Placing Task 2011 we will show the effectiveness and high accuracy relative to the state-of-the art solutions.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]

General Terms

Algorithms, Experimentation

Keywords

placing task, geotagging, hierarchical segmentation

ACM SIGSPATIAL LBSN '12 November 6, 2012. Redondo Beach, CA, USA

1. INTRODUCTION

Geo-coordinates are a form of metadata essential for organizing multimedia on the Web. Assigning geographical coordinates to shared content has become a popular activity for users in multimedia communities. Increasing numbers of capture devices such as cameras and smart phones automatically assign geo-coordinates to multimedia. Geo-coordinates enable users to find and retrieve data and allow for intuitive browsing and visualization. The majority of resources on the Web, especially videos, however, are not geo-tagged. Automatic methods for assigning geo-coordinates to video hold a large promise for improving access to video data in online multimedia communities.

The key contribution of this work is a framework for geo-tag prediction designed to exploit the relative advantages of textual and visual modalities. This approach is an extension of our previous work [14]. We will show that visual features alone show low correlation with locations and a purely visual approach achieves lower precision values than a purely tag-based approach. Indoor scenes, for example, are largely similar the world over, especially when images are represented in terms of low level features. However, in combination with a toponym lookup method that preselects videos of a possible area, even the weak visual information present in images improves geo-tagging performance—an effect that is demonstrated by our experiments. The paper is structured as follows. In the next section, we cover the related work. We introduce our approach using different modalities in section 4. The results are shown in section 5 and we finish with a conclusion summarizing our main findings.

2. RELATED WORK

Many approaches to geo-tagging based on textual gazetteers and visual analysis have been introduced previously. Kessler et al. [16] explain how existing standards can be combined to realize a gazetteer infrastructure allowing for bottom-up contribution as well as information exchange between different gazetteers. They show how to ensure the quality of user-contributed information and demonstrate how to improve querying and navigation using semantic-based information retrieval. Smart et al. [22] present a framework to access and integrate distributed gazetteer resources to build a meta-gazetteer, which generates augmented versions of place name information and combines different aspects of place name data from multiple gazetteer sources that refer to the same geographic place. At the end they employ

several similarity metrics to identify equivalent toponyms.

The approach of Hays et al. [12] is purely data-driven and their data is limited to a sub-set of Flickr images having only geographic tags. They find visual nearest neighbours to a single image based on low-level visual image descriptors and propagate the geo-location of the GPS-tagged neighbours. The approach by Hays et al. serves as a very general means for exploring similarities between images. By itself, it provided very limiting accuracy. Working with object retrieval methods, several authors [21] [5] build visual vocabularies which are usually created by clustering the descriptor vectors of local visual features such as SIFT.

Crandall et al. [10] propose a system to place images to a world map in combination with textual and visual information, trained with a dataset of about 35 million images collected from Flickr. They improve the ability to estimate the location of the photo using visual and time stamp features, compared to using just textual features. They build a binary classifier model for each of the ten landmarks of the city where the photograph was taken. Each photograph is represented by a feature vector consisting of vector-quantized SIFT features, which capture visual image properties, and text features extracted from the textual keyword tags.

The 2010 and 2011 MediaEval Placing tasks provided a common platform to evaluate different geo-tagging approaches on a corpus of randomly selected consumer-produced videos. Friedland et al. [9] addressed the case where the training data set is sparse and explored the possibility of using the test data set to improve the quality of the training database. They proposed a graphical model framework, posed the problem of geo-tagging as one of inference over this graph. Penatti et al. [19] introduced a visual-based geo-coding approach using a dictionary of scenes. The feature space spanned by such a model has the property of having one dimension for each semantic concept. The strategy of Sevillano et al. [20] based on extracting and expanding the geographical information contained in the textual metadata using Wikipedia as a gazetteer. If the input video contains no location names in their textual metadata the process based on a purely visual retrieval approach using the colour and edge directivity descriptor and the edge histogram descriptor.

3. PLACING TASK

Our experiments were conducted under the specifications of the 2011 Placing Task which is part of the MediaEval benchmarking initiative, that requires assigning geographical coordinates (latitude and longitude) to each provided test video. Here ,we can make use of metadata and audio and visual features as well as external resources, depending on the run. During the first year in 2010, there were no restrictions on what data or technique be used. The 2011 task encourage innovation in situations that reflected the constraints of realistic scenarios. For example, one run was required that used only the visual/audio content of the video for placing, which reflects the situation of needing to locate a video which has not yet tagged with any textual metadata.

3.1 Data sets

The MediaEval Placing Task 2011 required participants to predict geographical coordinates (latitude and longitude) for each provided test video. In order to achieve this goal, the participants can make use of metadata (e. g. title, description, tags, comments, etc.) and audio and visual features as well as external resources like gazetteers.

The training data provided for this task costists of 10,216 videos and 3.2 million images gathered from Flickr[1], uniformly sampled from all over the world, distributable under Creative Commons licenses. The test data is a separate set of 5,347 Flickr videos. The metadata for all videos and images includes any available metadata—tags, title, description and in same cases uploader information including contacts, favorites, gender and home location.

Figure 1 shows the sparse nature of the provided training data set for a map section of Europe. Each black dot represents one location of the training set and the red ones indicate the position of the each video sequence of the test data. The human eye traces out the Europe continent, but other areas of the world are less well covered than this map selection. This highlights the challenge of being able to locate videos in areas where there may not be many training examples.

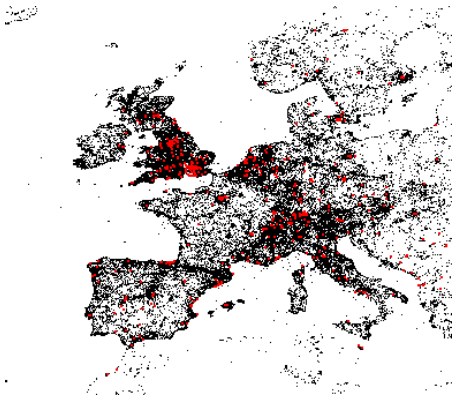


Figure 1: Sparse nature of the training set (black dots) in Europe. Red dots shows the location of the test examples.

Only those Flickr videos had been collected its associated geo-coordinates were stored with the highest accuracy level. The accuracy attribute encodes at which zoom level the uploader used when placing the video on a map. There are 16 zoom resp. accuracy levels (e. g. , 3 - country level, 6 - region level, 12 - city level, 16 - street level). All provided Flickr photos have at least region level accuracy.

The provided videos were accompanied by extracted key frames, which have been extracted every four seconds using FFmpeg[2]. For each of these key frames and for each Flickr image provided for training purposes, nine visual features were extracted using the open source library LIRE [17] with the default parameter settings.

- *Colour and Edge Directivity Descriptor (CEDD)* [8] combines color and texture information in a histogram.
- *Gabor Descriptor (GD)* [11] is a linear filter using frequency and orientation representations for edge detection.
- *Scalable color descriptor (SCD)* [6] uses vector wavelet coefficients of color images.

[1] `http://www.flickr.com/`
[2] `http://ffmpeg.org/`

- *Auto colour correlogram (ACC)* [13] extracts the spatial correlation of colors.
- *Tamura texture descriptor (TD)* [23] extracts histograms of low dimensional texture characteristics.
- *Edge histogram descriptor(EHD)* [18] extracts the distribution of 5 types of edges in each sub-image of 4×4 non-overlapping blocks.
- *Colour layout descriptor (CLD)* [18] is designed to capture the spatial distribution of color in an image.

3.2 Evaluation

The performance of each technique is evaluated using the geodesic distance between the ground truth coordinates and those of the prediction. To take into account the geographic nature of the task, the Haversine distance was used. This measure is calculated thus:

$$d = 2 \cdot r \cdot \arcsin\left(\sqrt{h}\right) \quad (1)$$

$$h = sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + cos(\phi_1)cos(\phi_2)sin^2\left(\frac{\psi_2 - \psi_1}{2}\right) \quad (2)$$

where d is the distance between points 1 and 2 represented as latitude (ϕ_1, ϕ_2) and longitude (ψ_1, ψ_2) and r is the radius of the Earth (the WGS-84 standard value of 6378.137km is used).

The following results should be considered with the following points in mind:

- The scope of possible video placement is considered to be the entire planet.
- This implies that the maximum possible distance between any two points is half the equatorial circumference, which is 20,037.51km according to WGS-84 standard. This provides an upper bound to any distance error. However, this can be improved by assuming a trivial video placing approach that assigns a test video the location of a randomly chosen training video. This would then provide an average upper bound distance of 12,249km using the 2011 training and test data.

Each judgement from a system was evaluated and grouped according to how close it was to the ground truth with respect to increasing distance: 1 km, 10 km, 20 km, 50 km, 100 km, 200 km, 500 km, 1,000 km, 2,000 km, 5,000 km, 10,000 km and 20,000 km.

4. FRAMEWORK

The participants in the Placing Task 2011 were allowed to use image/video metadata, external resources like gazetteers, audio and visual features in condition of the submitted run. Our proposed framework assigns geo-tags for Flickr videos based on their textual metadata and visual content in a hierarchical manner and includes several methods that are combined as depicted in figure 2. The first step is the pre-classification of these videos into possible regions on the map using the meridians and parallels. The key aspect to build these regions is the spatial segmentation of the geo-tagged database which generates visual and textual prototypes for each segment. The generation of segment prototypes are described in section 4.2.1 and 4.2.2. The national borders detection extracts toponyms and uses gazetteers to increase the effectiveness of our proposed approach. Finally, the probabilistic model superimposed all hierarchy levels and leads to the most similar image, based on the fact that there is a higher probability of two images taken at the same place. We choose this hierarchical way in order to reduce computational cost, since not all data of our database need not to compute for each training sample.

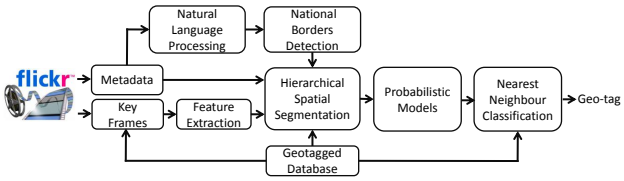


Figure 2: Textual and visual features are used in a hierarchical framework to predict the most likely location.

4.1 Hierarchical Spatial Segmentation

We tackle this geo-referencing problem with an classification approach in a hierarchical manner. Therefor, the world map is iteratively divided into segments of different sizes. The spatial segments of each hierarchy level is here considered as classes for our probabilistic model. Whereas the granularity is increased in lower hierarchy levels. So our classifiers are iteratively applied to classify video sequences to spatial locations becoming continual finer. These hierarchical segments are generated in two ways: querying gazetteers for toponyms and static segmenting with spatial grids different sizes.

4.1.1 National Borders Detection

In general, textual information (such as the provided metadata of the uploader) is a valuable source of information regarding the multimedia resource it is associated to. The *national borders method* extracts the geographical national borders using the toponyms extracted from the metadata which are used for looking up the geo-coordinates. For this purpose, the textual labelling is extracted from the video (e. g. description, title, and keywords) to collect all information about the possible location. Then, non-English metadata is handled by detecting the language and translating into English sentence by sentence. The translation is carried out using Google Translate [1], a free statistics-based machine translation web service. The translated metadata of the video to be geo-tagged is analysed by natural language processing (NLP) in order to extract nouns and noun phrases. For this task we use OpenNLP [7], a homogeneous package based on a machine learning approach that uses maximum entropy. NLP returns a huge list of candidates often including location information. Each item in the list is coarsely filtered using GeoNames [2]. The GeoNames database contains over 10 million geographical names corresponding to over 7.5 million unique features and provides a web-based search engine which returns a list of entries ordered by relevance. Next, we query Wikipedia [3] with each toponym candidate and examine the articles returned. The Examination involves parsing the Wikipedia article to determine whether it contains geo-coordinates. We take the presence of such coordinates as evidence that the toponym candi-

date is indeed a word associated with a place. If a candidate fails to return any Wikipedia articles, it is discarded. The Wikipedia filter constitutes a simple yet effective method for eliminating common nouns from the toponym candidate list.

The next step serves to eliminate *geographical ambiguity* among the toponym candidates. With the help of GeoNames, we create a rank sum $R(c_i)$ of each of the M possible countries c_i in which the place designated by all N toponym candidates could be located. The most likely country has the highest rank sum:

$$c_{detected} = argmax \begin{pmatrix} \sum_{j=0}^{N-1} R_j(c_0) \\ ... \\ \sum_{j=0}^{N-1} R_j(c_M) \end{pmatrix}.$$

The determination of a country is less ambiguous than that of a place or a city.

If there is no matching entity for any keyword in the metadata of the given video, this algorithm cannot detect any country borders and is analysing the whole world.

The geographical borders for a detected country are determined by querying the Google Maps API [4]. The resulting geographical borders supports the probabilistic models (sec. 4.2) in terms of preselecting likely spatial segments.

4.1.2 Spatial Segments of Different Granularity

The method of generating spatial segments divides the world map into areas of different granularities. The highest hierarchy level uses the *national borders detection* followed by a large grid of 360×180 segments according to the meridians and parallels of the world map. We also introduce a smaller grid of segments which spatial dimensions is halved to increase the accuracy and to minimise the computational cost. Each geo-tagged training image is assigned to its corresponding grid cell at the lowest level.

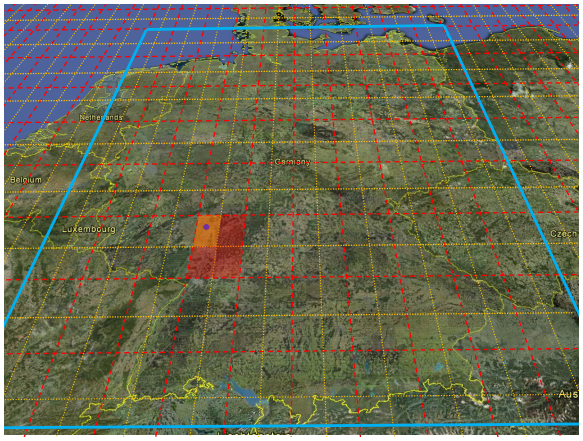


Figure 3: Visualization of hierarchical spatial segments for Central Europe: national borders detection for Germany (blue box), large segments (red boxes), small segments (orange boxes) and the geotagged items in the dataset (purple dot)

Figure 3 depicts our approach of hierarchical spatial segmentation for Central Europe after detecting Germany in the national borders detection.

4.2 Probabilistic Model

In this section the classification approaches are described that are used to determine the most likely spatial location at each hierarchy level. The both modalities—textual and visual—of each video sequence are separately geo-referenced at the most likely location, as shown in figure 4.

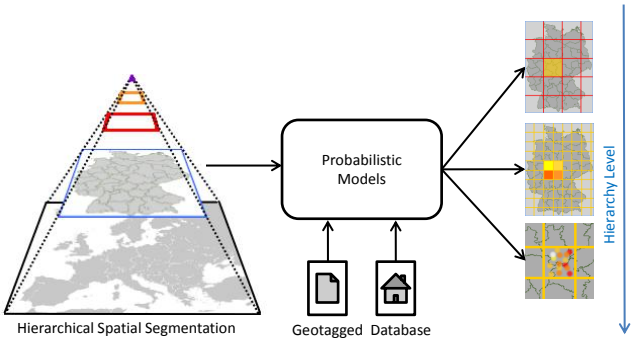


Figure 4: Geo-referencing at different levels using probabilistic models.

4.2.1 Textual Approach

The decision for spatial locations based on metadata can be regarded as classification of documents. A spatial location is either a specific area or a certain item, according to section 4.1. For applying a probabilistic classifier we treat the spatial locations l as classes. The data basis are geotagged images and videos with associated metadata from the training set assigned to the spatial locations. The vocabulary V of the spatial locations includes tags and words from the titles and descriptions. So each spatial segment incorporates all term occurences of its associated images from the training database. Each term from the vocabulary is stemmed using Porter stemmer algorithm[3], once stop words and digits were removed. For classifying the test video sequences d into locations l, their terms t are used in a probabilistic multinomial Bag-of-Words approach. So each sequence is iteratively assigned to the most likely spatial segment, according to the hierarchical segmentation:

$$l_{ml} = arg \max_{l \in L} P(d|l),$$

where $P(d|l)$ is the conditional probability that reflects the video sequence belonging to a certain location. This probability is defined by the term-location probability:

$$P(d|l) = P(< t_1, \ldots, t_{n_d} > |l),$$

where n_d is the number of terms in the video's metadata. Assuming the statistically independent of the term occurrence, the video-location probability is simplified to a multiplication of term-location probabilities:

$$P(d|l) = \prod_{k=1}^{n_d} P(t_k|l).$$

[3]http://tartarus.org/ martin/PorterStemmer/index.html

The use of logarithms replaces the multiplication by summation and preserves for floating point underflows:

$$\log(P(d|l)) = \sum_{k=1}^{V} N_{t_k,d} \cdot \log(P(t_k|l), \quad (3)$$

where $N_{t_k,d}$ is term frequency of term t_k in the metadata of video d. The term-location-distribution is estimated with the following formula that is smoothed by adding-one—which simply adds one to each count:

$$P(t|l) = \frac{N_{t,l} + 1}{\sum_{t' \in V} \left(N_{t',l} + 1 \right)}, \quad (4)$$

where $N_{t,l}$ is the term frequency of term t in a spatial segment l. The smoothing is necessary to have a probability value higher than zero for all terms t in all locations. These above formulas describe our probabilistic model when using a multinomial distribution with term frequency (tf) weighting. In latter studies we experiment with different weights, such as:

- Term frequency (TF).
- Term frequency-inverse document frequency (TF-IDF). The $N_{t_k,d}$ in Eq. 3 and $N_{t,l}$ in Eq. 4 are replaced by the tf-idf scores.
- Term occurrence (TO). The $N_{t_k,d}$ in Eq. 3 and $N_{t,l}$ in Eq. 4 are replaced by scores that indicates presence (1) or absense (0).

So each model generates the most likely location for each test video sequence at the given granularity within the hierarchy.

4.2.2 Visual Approach

This approach uses different visual features extracted from the Placing Task 2012 data base containing 3.2 million geotagged images and video sequences, respectively their key frames, to predict a location. Their visual content is described by all provided descriptors which covers a wide spectrum of descriptions of colour and texture within images. These image descriptions are pooled for each spatial segment in the different hierarchy level using the mean value of each descriptor. A k-d tree containing all appropriate segments is built for each descriptor and in each hierarchy level. This k-d tree has the advantage that the following search for nearest neighbour is speeded up because not all data needed to be computed. Following, the segment with the lowest distance becomes the most likely location at a given level of granularity. So, this method determines iteratively the most visually similar spatial segment by calculating the Euclidean norm.

For the test videos we reduced the temporal dimensionality by using the associated key frames. Other norms did not achieve better results than the L^2 norm used for comparison, according to prior experiments [15].

4.2.3 Fusion of Textual and Visual Approaches

The methods for predicting the hierarchical segments described in previous sections can be combined in multiple ways to synergise. The fusion can be done in the following way:

- Parallel mode (sum rule): The confidence scores of the textual and visual approaches are brought to the same scale and then combined using summation.
- Serial mode: Textual approach is used first for predicting, in case of absence of metadata the visual approach is applied.
- Serial mode (of hierarchies): The results of textual and visual approaches of different hierarchy levels are combined. Here the segments of higher hierarchy levels are predicted with the textual approach, while the spatial segment within the lowest (finest) hierarchy level is chosen using the visual approach.

5. EXPERIMENTS

In this section we describe the experimental setup for predicting the geographical coordinates where the respective video sequences were recorded. We run our experiments on the MediaEval 2011 placing task dataset which is described in details in section 3.1. The predicted locations of the 5,347 test video sequences are evaluated as described in section 3.2.

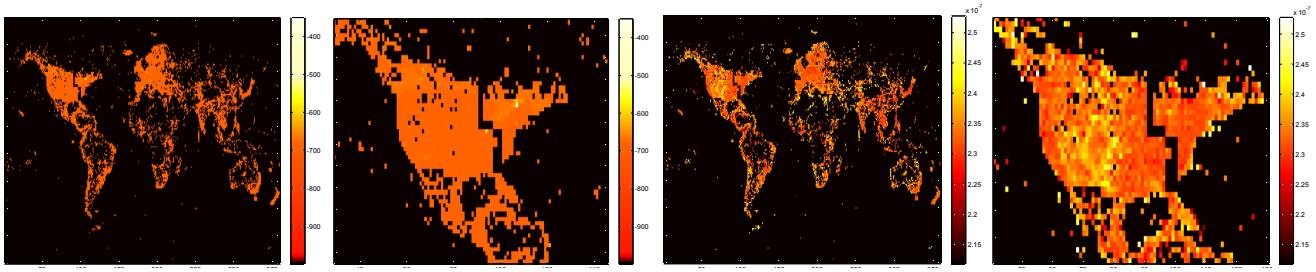
The results are discussed approach-wise in the following sections. Our results are compared to other state-of-the-art publications and to a baseline method that is based on randomness to show the statistical significance. For this purpose, each test video sequence is assigned the geographical coordinate of a randomly chosen training set item. This baseline method achieves an accuracy of about 12.3 % for an error of 1000 km.

5.1 Textual Approach

This section contains the results of the approach described in sec. 4.2.1. The three different weighting schemes of our model—term occurrence (TO), term frequency (TF), and term frequency-inverse document frequency (TF-IDF)—are compared against each other. Since the approach predicts iteratively the most likely spatial segments, we first evaluate the performance at the highest hierarchy level. The highest level corresponds the coarsest granularity within our segmentation of the world map, at this level the world is segmented according to the parallels and the meridians. The table 1 depicts the percentage of correct predicted spatial segments at coarsest scale. The textual model with TF-IDF weighting predicts the correct spatial segments for the half of the dataset. Considering the 10 most likely segments the location is correctly restricted in 66 % of all cases.

Table 1: Correct decision for spatial segments at the highest hierarchy level.

Top-N segments	TO	TF	TF-IDF
1	31.7 %	44.5 %	51.4 %
2	39.7 %	51.2 %	57.5 %
3	44.3 %	54.3 %	60,0 %
4	47.1 %	56.0 %	61.6 %
5	48.8 %	57.6 %	62.6 %
6	50.4 %	58.6 %	63.8 %
7	51.3 %	59.4 %	64.8 %
8	52.1 %	60.2 %	65.3 %
9	53.3 %	60.7 %	65.8 %
10	53.9 %	61.3 %	66.4 %



(a) world map (b) detail (c) world map (d) detail

Figure 5: Confidence scores of textual approach (a,b) and visual approach (c,d)

In general, the weighting with TF-IDF outperforms the other weightings. As expected, the accuracies on selected margin of errors are consistently higher for the TF-IDF weighting, as seen in table 2. We achieve a percentage of 56 % of correct predictions with city level (margin of error of 20 km).

Table 2: Accuracies on selected margin errors of the textual approach with different weightings.

margin of error	TO	TF	TF-IDF
1 km	13.9 %	15.0 %	19.4 %
10 km	30.0 %	39.5 %	46.8 %
20 km	35.8 %	46.7 %	56.0 %
50 km	41.5 %	54.6 %	64.0 %
100 km	45.7 %	58.7 %	66.8 %
200 km	50.7 %	62.4 %	71.0 %
500 km	57.3 %	62.4 %	74.8 %
1,000 km	63.3 %	70.9 %	78.2 %
2,000 km	72.0 %	75.1 %	82.5 %
5,000 km	84.5 %	85.9 %	89.7 %
10,000 km	95.0 %	96.4 %	97.6 %
20,000 km	100 %	100 %	100 %

Thus, the TF-IDF decrease the score of terms that occur in multiple spatial segments, this fact positively affects the performance. The model with term occurrence (TO), where all terms are threaten equally, has the contrary effect.

5.2 Visual Approach

The results of the approach described in sec. 4.2.2 are shown in table 3. The table contains the results for each descriptor and two hierarchy levels. Since each descriptor is handled in a separated way, those will be separately evaluated for figure out the most geo-related visual feature. The label 'large' stands for the spatial level which segments are generated according to the meridians and parallels and the segments of level labelled with 'small' are halved in each dimension, respectively.

As seen in table 3 the scalable colour descriptor (SCD) consistently outperforms the other descriptors. Consequently, scalable colour is the most geo-related visual feature, whereas the prediction at finer level ('small') achieves more accurate results than at the coarser level 'large'. We expect our textual approach to perform better than our visual approach, what proves true. It should be noticed that our best visual model (SCD) achieves three times more accurate result than random baseline (12 % at 1,000 km).

5.3 Fusion

As described in section 4.2.3 the confidence score of both modalities are combined. Since our textual approach achieves very strong results, the combination with the visual approach results does not gain much.

The figure 5 shows the confidence scores of both modalities for an example video[4] depicting a formula one scene captured in Montreal, Canada. The confidence score is coded in colours as follows; very unlikely spatial segments are depicted in black colour, the colour gets lighter with increasing likelihood of the segments. The figure 5 (a,b) shows the confidence scores of the textual approach with TF-IDF weighting in a log-scale. As seen, the segments around Montreal are more likely than other areas in the world. The scores of the visual approach using scalable colour as feature is depicted in figure 5 (c,d), here are many likely regions in the world—this video sequence may have been recorded at any locations in the world, and only a restriction based on textual metadata reduce the number of possible candidates.

Figure 6 shows such a restriction; the TF-IDF text model predict the most likely segment at the higher hierarchy levels and the visual SCD model predicts locations within this segments. As shown, the previous example is correctly assigned to the city of Montreal, Canada. Here, the fusion of textual and visual methods is important to eliminate geographical ambiguities.

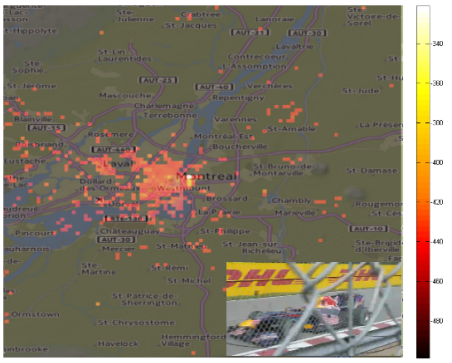


Figure 6: Confidence scores of the visual approach (SCD) restricted to be in the most likely spatial segment determined by the textual approach (TF-IDF).

Now, we compare our results against other state-of-the-

[4]http://www.flickr.com/photos/88878784@N00/4706267893

Table 3: Accuracies on selected margin errors (in km) of the visual approach with different descriptors.

Feature	Size	1	10	20	50	100	200	500	1,000	2,000	5,000	10,000	20,000
ACC	small	2.3 %	2.4 %	2.5 %	3.3 %	7.4 %	12.6 %	19.9 %	31.6 %	43.6 %	59.7 %	90.7 %	100 %
	large	2.3 %	2.4 %	2.5 %	3.3 %	7.4 %	12.6 %	19.9 %	31.6 %	43.6 %	59.7 %	90.7 %	100 %
CEDD	small	3.2 %	3.2 %	3.4 %	5.1 %	7.3 %	11.7 %	22.1 %	29.8 %	44.5 %	62.9 %	91 %	100 %
	large	2.3 %	2.4 %	2.5 %	3.3 %	7.4 %	12.6 %	19.9 %	31.6 %	43.6 %	59.7 %	90.7 %	100 %
CLD	small	1.2 %	1.3 %	1.4 %	2.2 %	5.9 %	11.9 %	18.6 %	28.5 %	45.2 %	60.9 %	90.3 %	99.4 %
	large	2.3 %	2.4 %	2.5 %	3.3 %	7.4 %	12.6 %	19.9 %	31.6 %	43.6 %	59.7 %	90.7 %	100 %
EHD	small	1.8 %	2 %	2.2 %	3.1 %	5.2 %	12 %	20 %	30.2 %	47.3 %	62.5 %	90.7 %	100 %
	large	2.3 %	2.4 %	2.5 %	3.3 %	7.4 %	12.6 %	19.9 %	31.6 %	43.6 %	59.7 %	90.7 %	100 %
GD	small	1.2 %	1.3 %	1.3 %	2.3 %	4 %	7.1 %	12.5 %	24.2 %	37 %	65.2 %	89.8 %	100 %
	large	2.3 %	2.4 %	2.5 %	3.3 %	7.4 %	12.6 %	19.9 %	31.6 %	43.6 %	59.7 %	90.7 %	100 %
TD	small	0.7 %	0.7 %	0.7 %	1.4 %	4.7 %	9.6 %	15.3 %	21.6 %	37 %	55.9 %	89.7 %	100 %
	large	2.3 %	2.4 %	2.5 %	3.3 %	7.4 %	12.6 %	19.9 %	31.6 %	43.6 %	59.7 %	90.7 %	100 %
SCD	small	5.4 %	5.6 %	5.8 %	6.5 %	8.6 %	13.8 %	24.2 %	34.9 %	50.2 %	63.3 %	90.5 %	100 %
	large	2.3 %	2.4 %	2.5 %	3.3 %	7.4 %	12.6 %	19.9 %	31.6 %	43.6 %	59.7 %	90.7 %	100 %

art publications and against a random baseline. The figure 7 show results plotted against the geographical margin of error. The blue solid line (TUB) shows the results of our proposed approach with the textual TF-IDF model and the visual SCD model in serial mode. The red dashes line (QMUL) shows the results reported in Sevillano et al. [20], the results of the green dashed line (UNICAMP) are reported in Penatti et al. [19] and the purple dotted line visualise a random baseline. While the approach of Penatti et al. is purely data driven, it is significant worse than the other approaches—but we have shown in table 3 better results achieved with visual features only.

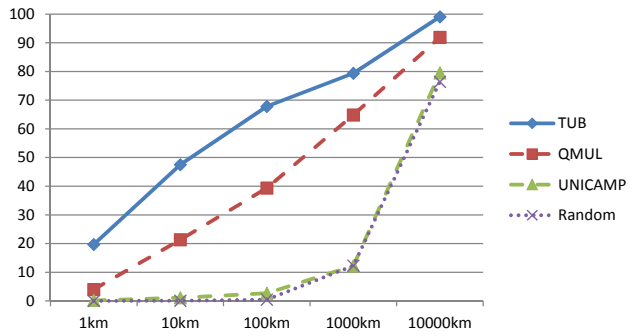


Figure 7: Accuracy plot against geographical margin of error: Comparison

As seen, our approach outperform the other methods, especial on smaller margin of errors. For a margin of error of 10 km, we achieve an accuracy of 47.5 % which doubles the accuracy of QMUL.

6. CONCLUSIONS AND FUTURE WORK

In this paper we presented a hierarchical approach for the automatic estimation of geo-tags in social media website such as Flickr. We presented a detailed analysis of textual and visual features using different spatial granularities and national borders detection. This external resources—using GeoNames and Wikipedia—are databases with still growing knowledge, therefore a training step is not needed. The fusion of textual and visual methods is important to eliminate geographical ambiguities. Finally, we have shown that our proposed approach retrieve a high accuracy relative to the state-of-the art solutions at the Placing Task 2011. We hereby showed that our framework is able to handle this geographically highly skewed distribution of Flickr media. We would like to point out that we are able to find a geolocation that is correctly located within a radius of 10 km for half of the test set. These results are encouraging and they leave a lot of potential for future work. We will improve our framework by using more distinctive visual descriptors (e. g. local features) and possibly object recognition algorithms, which can be applied to media items to predict locations accurately almost to the metre; a photograph depicting the Eiffel Tower, for instance, can be tagged precisely using external information, like images of the geo-tagged Wikipedia article.

7. ACKNOWLEDGMENTS

We would like to acknowledge the 2011 Placing Task of the MediaEval Multimedia Benchmark for providing the data used in this research. The research leading to these results has received funding from the European Community's FP7 under grant agreement number 261743 (NoE VideoSense).

8. REFERENCES

[1] http://translate.google.com.
[2] http://www.geonames.org.
[3] http://www.wikipedia.org.
[4] http://code.google.com/apis/maps/index.html.
[5] S. Agarwal, N. Snavely, I. Simon, S. Seitz, and R. Szeliski. Building rome in a day. In *Computer Vision, 2009 IEEE 12th International Conference on.*
[6] E. Albuz, E. Kocalar, and A. Khokhar. Scalable color image indexing and retrieval using vector wavelets. *Knowledge and Data Engineering, IEEE Transactions on*, 13(5):851–861, 2001.
[7] J. Baldridge. The OpenNLP Project. http://www.opennlp.com, 2005.
[8] S. Chatzichristofis and Y. Boutalis. Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. *Computer Vision Systems*, pages 312–322, 2008.

[9] J. Choi, G. Friedland, V. Ekambaram, and K. Ramchandran. Multimodal location estimation of consumer media: Dealing with sparse training data. In *proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2012), Melbourne, Australia*, 2012.

[10] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the World's Photos. In *Proceedings of the 18th international conference on World wide web*, pages 761–770. ACM, 2009.

[11] H. Feichtinger and T. Strohmer. *Gabor analysis and algorithms: Theory and applications*. Birkhauser, 1998.

[12] J. Hays and A. Efros. Im2gps: estimating geographic information from a single image. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. Ieee, 2008.

[13] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih. Image indexing using color correlograms. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 762–768. IEEE, 1997.

[14] P. Kelm, S. Schmiedeke, and T. Sikora. A hierarchical, multi-modal approach for placing videos on the map using millions of flickr photographs. In *ACM Multimedia 2011 (Workshop on Social and Behavioral Networked Media Access - SBNMA)*. ACM, Nov. 2011.

[15] P. Kelm, S. Schmiedeke, and T. Sikora. Multi-modal, multi-resource methods for placing Flickr videos on the map. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, New York, NY, USA, 2011. ACM.

[16] 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 International Conference on Advances in Geographic Information Systems*, pages 91–100. ACM, 2009.

[17] M. Lux and S. Chatzichristofis. Lire: lucene image retrieval: an extensible java cbir library. In *Proceeding of the 16th ACM international conference on Multimedia*, pages 1085–1088. ACM, 2008.

[18] B. Manjunath, J. Ohm, V. Vasudevan, and A. Yamada. Color and texture descriptors. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(6):703–715, 2001.

[19] 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 *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ICMR '12, pages 53:1–53:8, New York, NY, USA, 2012. ACM.

[20] X. Sevillano, T. Piatrik, K. Chandramouli, Q. Zhang, and E. Izquierdoy. Geo-tagging online videos using semantic expansion and visual analysis. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2012 13th International Workshop on*, pages 1–4. IEEE, 2012.

[21] I. Simon, N. Snavely, and S. Seitz. Scene summarization for online image collections. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[22] P. Smart, C. Jones, and F. Twaroch. Multi-source toponym data integration and mediation for a meta-gazetteer service. In *Geographic Information Science*, Lecture Notes in Computer Science.

[23] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *Systems, Man and Cybernetics, IEEE Transactions on*, 8(6):460–473, 1978.