A Novel Fusion Method for Integrating Multiple Modalities and Knowledge for Multimodal Location Estimation

Pascal Kelm[1], Sebastian Schmiedeke[1], Jaeyoung Choi[2],
Gerald Friedland[2], Venkatesan Ekambaram[3],
Kannan Ramchandran[3] and Thomas Sikora[1]

[1]Communication Systems Group, Technische Universität Berlin, Germany
[2]International Computer Science Institute, Berkeley, CA, USA
[3]Department of Electrical Engineering and Computer Sciences, UC Berkeley, Berkeley, CA, USA

[1]{kelm,schmiedeke,sikora}@nue.tu-berlin.de
[2]{jaeyoung,fractor}@icsi.berkeley.edu
[3]{venkyne,kannanr}@eecs.berkeley.edu

ABSTRACT

This article describes a novel fusion approach using multiple modalities and knowledge sources that improves the accuracy of multimodal location estimation algorithms. The problem of "multimodal location estimation" or "placing" involves associating geo-locations with consumer-produced multimedia data like videos or photos that have not been tagged using GPS. Our algorithm effectively integrates data from the visual and textual modalities with external geographical knowledge bases by building a hierarchical model that combines data-driven and semantic methods to group visual and textual features together within geographical regions. We evaluate our algorithm on the MediaEval 2010 Placing Task dataset and show that our system significantly outperforms other state-of-the-art approaches, successfully locating about 40 % of the videos to within a radius of 100 m.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Location Estimation

General Terms

geo-tagging, hierarchical segmentation, multimodal location estimation, centroid-based fusion

1. INTRODUCTION

Geo-coordinates are a form of metadata essential for organizing multimedia content on the Web, as location and time together form a physically guaranteed unique key. As a result, an increasing number of 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. However, given that only about 5 % of media are being geo-tagged [9], there exists a need for a method to automatically geo-tag the rest of the media available on the map, to improve multimedia organization, retrieval, semantic understanding, etc. The task of estimating the location of a media recording that lacks geo-location metadata has been referred to variously as "geo-tagging", "location estimation" or "placing". Location estimation helps provide geo-location where it is not usually available. For instance, except for specialized solutions, GPS is not available indoors or where there is no line of sight with satellites. Also, vacation videos and photos could be grouped even if geo-tags were not attached at the time of recording. As discussed in [13], this was one of the main motivations in the MediaEval benchmark[1].

The key contribution of this article is a fusion method for geo-tag prediction designed to exploit the different advantages of textual and visual modalities. We outperform previous approaches to multimodal location estimation and achieve a better performance than any that has previously been measured, locating 40 % of the videos to within a 100 m radius. Our experiments indicate that visual features alone show low correlation with locations and that a purely visual approach achieves lower precision values than a purely tag-based approach. However, in combination with a toponym lookup method that preselects videos of a possible area, even weak visual information from images improves geo-tagging performance.

This paper is structured as follows. In the next section, we cover related work; we introduce our approach to using different modalities in section 3; the results are described in section 4; and in section 5 we summarize our main findings.

2. RELATED WORK

Many approaches to geo-tagging based on textual gazetteers and visual analysis have been introduced previously. Previous work in the area of automatic geo-tagging of mul-

[1]http://www.multimediaeval.org

GeoMM'13, October 21, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2509230.2509238.

timedia based on tags has been mostly carried out on Flickr images. In [14], the geo-locations associated with specific Flickr tags are predicted using spatial distributions of tag use. A tag that is strongly concentrated in a specific location has a semantic relationship with that location.

The approach of Hays et al. [10] is purely data-driven, and their data is limited to a subset 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 of Hays et al. serves as a very general means for exploring similarities between images. By itself, it provided very limited accuracy.

The MediaEval Placing Task provided a common platform to evaluate different geo-tagging approaches on a corpus of randomly selected consumer-produced videos. In this paper, we evaluate our algorithm on the MediaEval 2010 Placing Task data set; in section 4, we compare our results with state-of-the-art approaches. One of many notable approaches was presented by Van Laere et al. [15] (Ghent) used combination of language models and similarity search to geo-tag the videos using their associated tags. Perea-Ortega et al. [11] (SINAI) and Ferrés et al. [8] (TALP) used text features in combination with gazetteers for the location estimation. If there was no gazetteer matching entity for all keywords in the metadata of a given video, Choi et al. [7] (ICSI) returned the geo-coordinate of the user's home location. Just one research team reported results using only visual content in 2010. Kelm et al. [12] (TUB) used visual features of the development set for training a multi-class Support Vector Machine (SVM) classifier with Radial Basis Function (RBF) kernel. Their best results were achieved by a hierarchical clustering with a diameter threshold of 100 km, which determined 317 classes for the SVM.

3. FRAMEWORK

The participants in the Placing Task were allowed to use image/video metadata, external resources like gazetteers, and audio and visual features. 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 1. The first step is the pre-classification of these videos into possible regions on the map using the meridians and parallels. The key to building 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 3.2.1 and 3.2.2. The national borders detection extracts toponyms and uses gazetteers to increase the effectiveness of our proposed approach. Finally, the probabilistic model superimposes 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 approach in order to reduce computational cost, as only a portion of the data in our database is needed to compute for each training sample. Our contribution in this paper is developing a framework that helps exploit the power of collective processing. We describe a centroid-based matching approach in section 3.3 to solve the problem of data sparsity.

3.1 Hierarchical Spatial Segmentation

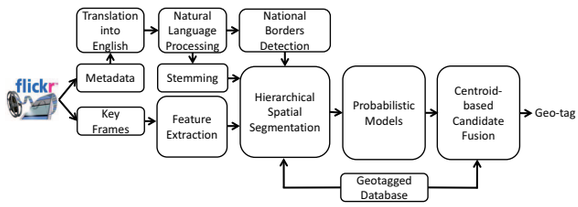


Figure 1: Textual and visual features are used in a hierarchical framework to predict the most likely location.

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

3.1.1 National Borders Detection

In general, textual information (such as the provided metadata of the uploader) is a valuable source of information about the multimedia resource it is associated with. The *national borders method* extracts geographical national borders using the toponyms extracted from the metadata, which are used for looking up the geo-coordinates. For this purpose, the textual label 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 statistically-based machine translation web service. The translated metadata of the video to be geo-tagged is analysed using natural language processing (NLP) to extract nouns and noun phrases. For this task we use OpenNLP [5], 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 candidate 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:

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

The determination of a country is less ambiguous than that of a landmark or a city. The geographical borders for a detected country are determined by querying the Google Maps API [4]. The resulting geographical borders support the probabilistic models (sec. 3.2) by preselecting likely spatial segments. If there is no matching entity for any keyword in the metadata of the given video, this algorithm cannot detect any country borders and analyses the whole world.

3.1.2 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.

3.2 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 two modalities—textual and visual—for each video sequence are separately geo-referenced to the most likely location.

3.2.1 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 3.1. To apply a probabilistic classifier, we treat the spatial locations l as classes. The base data are geo-tagged 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. Each spatial segment incorporates all terms related to its associated images from the training database. Each term from the vocabulary is stemmed using Porter stemmer algorithm[2], 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:

$$P(d|l) \sim P(l|d) \quad (2)$$

$$l_{ml} = arg \max_{l \in L} P(d|l), \quad (3)$$

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:

$$P(d|l) = P(< t_1, \ldots, t_{n_d} > |l), \quad (4)$$

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

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

$$P(d|l) = \prod_{k=1}^{n_d} P(t_k|l). \quad (5)$$

The use of logarithms replaces the multiplication by summation and prevents the underflow of floating points.

$$\log(P(d|l)) = \sum_{k=1}^{V} N_{t_k,d} \cdot \log(P(t_k|l), \quad (6)$$

where $N_{t_k,d}$ is the term frequency of the 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:

$$P\left(t|l\right) = \frac{N_{t,l} + 1}{\sum_{t' \in V} \left(N_{t',l} + 1\right)}, \quad (7)$$

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. The 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. 6 and $N_{t,l}$ in eq. 7 are replaced by the tf-idf scores.
- Term occurrence (TO). The $N_{t_k,d}$ in equation 6 and $N_{t,l}$ in equation 7 are replaced by scores that indicate presence (1) or absence (0).

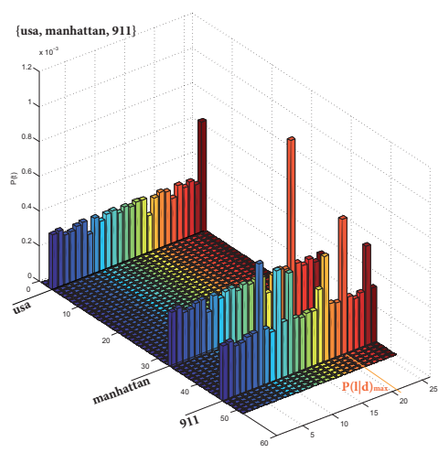


Figure 2: Term-location probabilities $P(t|l)$ of a video containing the terms: "usa", "manhattan" and "911".

So each model generates the most likely location for each test video sequence at the given granularity within the hierarchy. Figure 2 shows the term-location probabilities of

terms occurring in a video located at the Ground Zero, Mahattan, New York, USA. As shown, single terms do not indicate a specific or right location, only the combination maximizes the likelihood of the right spatial segment. Thus we can conclude that generic terms like "911" do help with specifying the location despite the fact that these terms do not have a geographical relation in the sense of being an entry in a gazetteer.

3.2.2 *Visual Approach*

This approach uses different visual features extracted from the Placing Task 2010 database to predict a video's location. The database contains 3.2 million geotagged images, video sequences and their respective key frames. The visual content of each video is described by the provided descriptors that cover a wide spectrum of descriptions of colour and texture within the keyframes. 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 that contains all the appropriate segments is built for each descriptor in each hierarchy level. This k-d tree has the advantage that the following search for nearest neighbour is speeded up because only a portion of data are needed to be computed. Then the segment with the lowest distance becomes the most likely location at the given level of granularity. 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 keyframes.

ïż£

3.3 Centroid-based Candidate Fusion

We are interested in addressing cases where the training dataset is sparse, thus we explored the possibility of using the test data set to aid the handling of sparsity in the training database. Choi et al. [6] proposed a graphical model framework and posed the problem of geo-tagging as one of inference over the graph, and showed that performance improvements can be achieved by smart processing of the test data set. The node potentials in this graph-based framework are further modeled as a product of the distributions, given each tag individually.

In the following section, we describe a centroid-based candidate fusion to solve the problem of data sparsity and to enhance the distributions of single candidates in a multimodal manner. The framework also facilitates the fusion of textual and visual features that can further improve the localization performance. One of the biggest problems in the fusion of multimodal features is the different range of features from each of the domains that are used.

Our centroid-based candidate fusion approach is based on the sum rule in decision fusion. Since our textual location model produces logarithmic confidence scores (see eq. 6) for location candidate, these scores have to be exponentiated. Subsequently, these scores are used to generate normalized weights for the candidate fusion, according to equ. 8:

$$w_n = \frac{P(l_n|d)}{\sum_{i=1}^{N} P(l_i|d)}, \tag{8}$$

where w_n is the weight of the nth candidate of a test video d. The equation implies that the sum of weights is 1. Then, the candidate location $GPS(\cdot)$ are weighted combined using the sum rule, assumming that the most likely location has the highest weight. The location centroid $\mathbf{x}^{v|t}$ arising from all visual(v) or textual(t) candidates is calculated as follows:

$$\mathbf{x}^{v|t} = \sum_{n=1}^{N} w_n \cdot GPS(l_n). \tag{9}$$

This forms the most likely location for a given video, whereas the centroid does not need to be an existing item within the training data. Thus the sparsity is not that problematic. Then, we determine the location for a specific feature. Since we want to include several different features from different modalities in our framework, a confidence score for each calculated centroid is needed for the subsequent multimodal fusion. Here, we choose the standard deviation as inversely proportional weights. The more a feature correlates with a specific spatial location, the closer the likely candidates are located, which implies a small deviation and therefore a high-valued weight. The calculation of the spatial deviation is shown in the following equation:

$$\sigma^{v|t} = \sqrt{\sum_{n=1}^{N} \left(GPS\left(l_n\right) - GPS\left(\mathbf{x}^{v|t}\right)\right)^2 \cdot w_n} \tag{10}$$

Using these formulas, a centroid is determined for each feature. The final decision for the video location X is specified as shown in fig. 3 using the multimodal fusion in eq. 11:

$$\mathbf{X} = w^t \cdot \mathbf{X}^t + w^v \cdot \mathbf{X}^v, \tag{11}$$

where $w^{v|t}$ are calculated according to equation 8, but with $\frac{1}{\sigma}$ instead of $P(l_i|d)$.

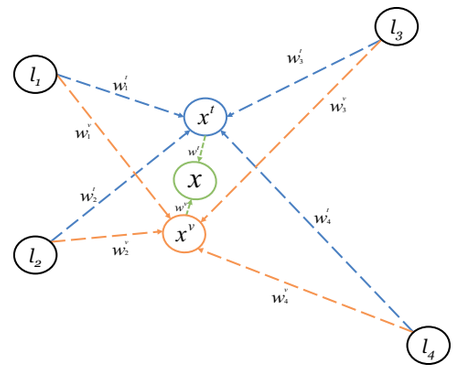


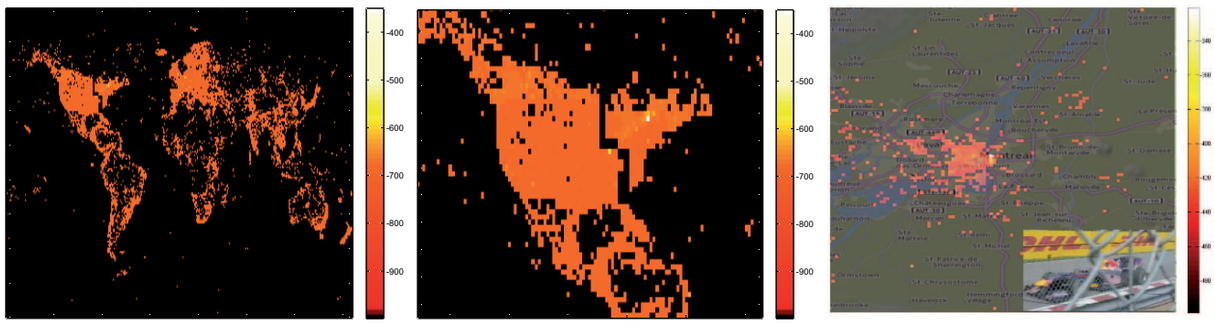
Figure 3: Illustration of fusion of textual and visual candidates.

4. 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 2010 placing task dataset.

We compare our results to other state-of-the-art systems that were tested on the same dataset. All results are also compared to a random baseline. For this purpose, each test video sequence is assigned the geographical coordinate of a randomly chosen training set item. This random baseline achieves an accuracy of about 12.3 % for an error of 1000 km.

This section contains the results of the approach described in section 3. Since we use a method for national borders de-



(a) World map (b) North America (c) Candidates within most likely spatial segment

Figure 4: Confidence scores (in log scale) of textual approach (a, b) and the centroid-based candidate fusion within the most similar block (c).

tection which restricts candidates located in a specific area, we first evaluate the performance of our hierarchical probabilistic model. Although this dataset is surprisingly well tagged, a purely gazetteer-based approach cannot identify all video locations, because 10 % of the test video does not contain location related metadata.

We introduced three different weighting schemes for our model–term occurrence (TO), term frequency (TF), and term frequency-inverse document frequency (TF-IDF)–which are now compared against each other. Table 1 shows the percentage of correctly predicted spatial candidates using restrictions made by our national borders detection.

Table 1: Accuracies for selected margins of error for different textual weightings without previous national borders detection.

margin of error	TO	TF	TF-IDF
1 km	38.5 %	50.9 %	66.5 %
10 km	55.9 %	59.9 %	74.4 %
20 km	62.6 %	63.8 %	77.4 %
50 km	72.5 %	73.6 %	82.6 %
100 km	82.0 %	79.4 %	86.7 %
200 km	94.3 %	93.8 %	95.3 %
500 km	100 %	100 %	100 %

The weighting with TF-IDF outperforms both of the other weighting schemes, as the TF-IDF weighting correctly decreases the score of terms which occur in multiple spatial segments. This can be observed by comparing to the pure TF model. In contrast, the model with term occurrence (TO), where all terms are treated equally, seems to have the contrary effect, thus performing much worse. Summarizing this experiment, we conclude that even without the use of gazetteers, we are able to predict the location of 74.4 % of the test videos within a radius of 10 km.

The next set of experiments shows that the accuracies increase when we use the pre-selection of spatial segments made by the national borders detection, as seen in table 2.

Within a radius of 10 km, we achieve 86.8 % in correct location predictions. Figure 4 shows the confidence scores for candidates for an example video[3] depicting a Formula

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

Table 2: Accuracies for selected margins of error for different textual weightings with previous national borders detection

margin of error	TO	TF	TF-IDF
1 km	70.0 %	74.4 %	78.3 %
10 km	84.1 %	84.6 %	86.8 %
20 km	87.7 %	87.7 %	89.7 %
50 km	92.5 %	92.6 %	93.5 %
100 km	98.8 %	95.3 %	96.2 %
200 km	100 %	98.4 %	98.8 %
500 km	100 %	100 %	100 %

One scene captured in Montreal, Canada. The confidence score is colour-coded as follows: very unlikely spatial segments are depicted in black, and the colour gets lighter as the likelihood of the segments increases. Figure 4 shows the confidence scores of the probabilistic approach in a log-scale for spatial segments in the world (a), for spatial segments in North America (b) and for candidates in the most likely segment (c). As seen, the segments around Montreal, Canada are more likely than other areas in the world.

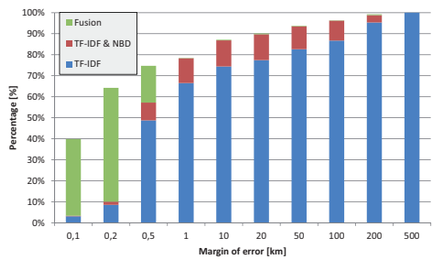


Figure 5: Accuracy achieved by single methods within the hierarchy for selected margins of error.

The candidates of both modalities are combined using our centroid-based fusion as described in section 3.3. The videos of the Placing Task dataset are well tagged, the textual model produced strong candidates and the combination with the visual candidates effect a location gain in small

scale. In general, the fusion of several candidates of both modalities is important to eliminate geographical ambiguities. As depicted in figure 5, the centroid-based fusion improves the results, especially on smaller margins of error, overcoming the sparse nature of the dataset. The strongest gain is achieved by the gazetteer-based national border detection, which eliminates the geographical ambiguity in conjunction with the probabilistic models.

Finally, we compare our fusion results against other state-of-the-art results. Figure 6 shows results plotted against the geographical margin of error of other systems. The green solid line shows the results of our proposed approach with centroid-based multimodal candidate fusion. The results of the other approaches are taken from MediaEval 2010 Placing Task participants. The blue dashed line (TUB) shows the results of a multimodal approach reported in Kelm et al. [12]. The results represented by the purple line (Ghent) are reported in van Laere et al. [15], and the red line represents the results of Perea-Ortega et al. (SINAI) [11]. The results of Choi et al. (ICSI) [7] and Ferrés et al. (TALP) [8] are depicted as a blue line and a brown line, respectively. As can be seen, our approach outperforms the other approaches. Our fusion approach, which combines gazetteers and probabilistic models, is significantly better at eliminating geographical ambiguity than competing approaches. For a margin of error of 1 km, we achieve an accuracy of 78.5 % which doubles the accuracy of Ghent, who uses a purely tag-driven approach, and outperforms the other gazetteer-based approaches.

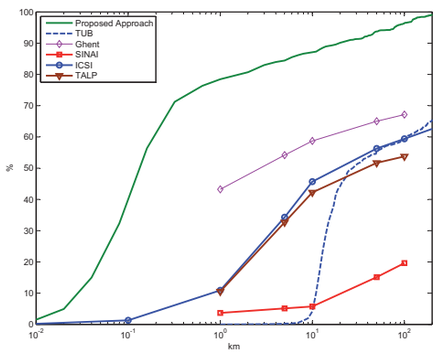


Figure 6: Accuracy plot against geographical margin of error, comparing MediaEval 2010 Placing Tasks participants.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a novel fusion approach that improves automatic estimation of geo-tags for social media videos. Our approach locates 40 % of the videos to within 100 m of the ground truth location, and 78.5 % of the videos to within 1 km.

We presented a detailed analysis of this fusion approach, which uses textual and visual features at different spatial granularities, and described the effect of national borders detection. We showed that our fusion approach is effective at eliminating geographical ambiguities. Given that only about 5% of Internet videos are geo-tagged [9], there is a pressing need to develop effective estimation models. According to our experiments, our approach significantly increases the number of geo-tagged media at a granularity of 100 m.

We plan to work on adding acoustic location estimation.

6. ACKNOWLEDGMENTS

We would like to acknowledge the Placing Task organizer of the MediaEval Multimedia Benchmark for providing the data used in this research.

7. 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] J. Baldridge. The OpenNLP Project. http://www.opennlp.com, 2005.
[6] J. Choi, G. Friedland, V. Ekambaram, and K. Ramchandran. Multimodal location estimation of consumer media: Dealing with sparse training data. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pages 43–48, July.
[7] J. Choi, A. Janin, and G. Friedland. The 2010 ICSI Video Location Estimation System. In *Working Notes of the MediaEval 2010 Workshop*, 2010.
[8] D. Ferrés and H. Rodríguez. Talp at mediaeval 2010 placing task: Geographical focus detection of flickr textual annotations. *Proceedings of MediaEval*, 2010.
[9] G. Friedland, O. Vinyals, and T. Darrell. Multimodal Location Estimation. In *Proceedings of ACM Multimedia*, pages 1245–1251, 2010.
[10] J. Hays and A. A. Efros. IM2GPS: estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, pages 1–8. IEEE, June 2008.
[11] L. A. U.-L. M. G.-V. José M. Perea-Ortega, Miguel A. García-Cumbreras. In *Working Notes Proceedings of the MediaEval 2010 Workshop*, Pisa, Italy, 2010.
[12] P. Kelm, S. Schmiedeke, and T. Sikora. Multi-modal, multi-resource methods for placing flickr videos on the map. In *ACM International Conference on Multimedia Retrieval (ICMR)*, page 8, Apr. 2011.
[13] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and G. J. F. Jones. Automatic tagging and geotagging in video collections and communities. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 51:1–51:8, New York, NY, USA, 2011. ACM.
[14] T. Rattenbury and M. Naaman. Methods for extracting place semantics from flickr tags. *ACM Trans. Web*, 3(1):1:1–1:30, Jan. 2009.
[15] O. Van Laere, S. Schockaert, and B. Dhoedt. Ghent University at the 2010 Placing Task. In *Proceedings of MediaEval*, October 2010.