Video2GPS: Geotagging using collaborative systems, textual and visual features

MediaEval 2010 Placing Task

Pascal Kelm, Sebastian Schmiedeke, Thomas Sikora
Communication Systems Group
Technische Universität Berlin
Germany
{kelm,schmiedeke,sikora}@nue.tu-berlin.de

ABSTRACT

Assigning geographical coordinates to shared content has become a popular activity on the Web, but nevertheless there are still huge amounts of media data without any geographical tags. Our approach enables these media data to be geotagged with the help of recently tagged media and knowledge-based collaborative systems. It includes three different methods–querying collaborative systems, document indexing, and classification based on visual features –which are combined to estimate geographical regions. Our approach can be applied under restrictions encountered in real applications. We give five experimental results for the MediaEval 2010 placing task collection.

General Terms

geolocalization, collaborative systems, probabilistic latent semantic analysis, support vector machine, mpeg-7

1. INTRODUCTION

There are millions images and videos on Flickr alone that are placed on the world map either manually by the uploader or automatically by GPS devices. So it is already possible to browse for Flickr media just by clicking the map. The majority of media data, however, is not geotagged. Recent works in the area of information retrieval and computer vision address the estimation of the geographical location of media items. The approach of Hays and Efros [1] is purely data-driven. They find visual nearest neighbours to a single image and propagate the geolocation of the GPS-tagged neighbours. Their results are quite good, but we show that the use of available metadata dramatically increases the precision. A pure text-driven approach for placing Flickr images on a map is presented by Serdyukov et al [4]. They divide the world map into cells of different size (from 1 km up to 100 km). A bag-of-words approach is used to model the location distribution per tag and GeoNames is used to acquire further knowledge whether a tag is location-specific. Their results are better than those achieved by methods relying only on visual data in [1]. We present an approach that fuses visual and textual methods with collaborative systems in order to achieve higher accuracy.

MediaEval 2010 Workshop, October 24, 2010, Pisa, Italy
.

2. PROPOSED FRAMEWORK

Our proposed framework includes three approaches using textual and visual information of shared media. The metadata (e.g. description, tags, title, etc.) are only available in different languages, which complicates the natural language processing (NLP). For this reason we detect the language and we translate the text into English using the web service Google Translate[1]. This (translated) textual information is processed by two independent approaches–probabilistic latent semantic analysis (pLSA) and collaborative systems. The visual approach uses colour and edge features [3] to train a support vector machine (SVM). The methods are combined as depicted in figure 1.

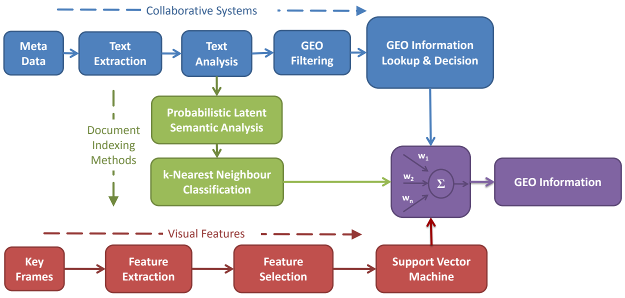

Figure 1: System overview

2.1 Collaborative Systems

We use collaborative systems for an effective knowledge database access to bridge the semantic gap.The translated metadata of the video to be geotagged is analysed by NLP[2] in order to extract nouns. Nouns with geographical context are filtered with help of Wikipedia[3]. The resulting words are analysed by GeoNames[4] with respect to their geographical ambiguity and their category. The ambiguity of nouns is reduced by finding the largest intersection of the basic category, i.e. country. Now the geographical ambiguity is reduced to places with that category (i.e. places in that country). The final decision is made by a modified rank

[1] http://translate.google.com
[2] http://www.opennlp.com
[3] http://www.wikipedia.org
[4] http://www.geonames.org

Table 1: 5 fusion experiments with accuracy on selected margin of errors

experiment	collaborative systems	document indexing method	visual features	5km	50km	100km
`gupnk1`	✓& uploader	translated nouns, k=1	*X*	32.92%	56.69%	60.16%
`gupnk7`	✓& uploader	translated nouns, k=7	*X*	32.92%	56.71%	60.16%
`gpnk7`	✓	translated nouns, k=7	*X*	20.04%	32.63%	35.02%
`gpkk7`	✓	only key words, k=7	*X*	20.04%	32.61%	35.04%
`guvswv`	✓& uploader	*X*	✓	32.92%	56.73%	60.46%

sum in which higher-level categories (e.g. city) and popular location are ranked higher. The resulting geotag is obtained from the highest category (i.e. place) of the decision.

2.2 Document Indexing Methods

In order to compute densities for textual and visual features, the world map is segmented 360×180 geoblocks. The videos of the training set are assigned to these blocks according to their geotag. An occurrence matrix $n(d_i, w_j)$ of videos d_i and their metadata terms w_j is generated. According to the different trials (table 1) the terms are either the extracted (English) nouns or pure tags. Based on the occurrence matrix, pLSA [2] models the probability of each co-occurrence of terms as a mixture of conditionally independent multinomial distributions:

$$P(d_i, w_j) = \sum_l P(z_l) P(d_i|z_l) P(w_j|z_l).$$

For each video from the training set the probability vector $P(z|d_i)$ of L latent topics z_l is calculated, where the number of latent topics M is set equal to the number of geoblocks:

$$P(z_l|d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j) P(z_l|d_i, w_j)}{n(d_i)}.$$

k-nearest neighbour algorithm is used to assign geographical tags to the test videos by comparing the probability vector $P(z|d_{test})$ with the corresponding one from the training set. So the test video is assigned the geotag of the geoblock that contains most of the k most similar training videos.

2.3 Visual Features

We extract nine features from each key frame of the video sequences and we reduce them to five features using a forward features selection algorithm in order to decrease the dimensionality. These five features (Color and Edge Directivity Descriptor, Scalable Color, Edge Histogram, Fuzzy Color and Texture Histogram, Color Layout) are used to train a SVM classifier that produces probabilities for each geoblock. The geoblocks are generated as described in section 2.2; figure 2 shows an example probability map. The geoblock with the highest probability is assigned to the key frame. The probabilities of the key frames of one video form the basis of a weighted voting to obtain a single decision.

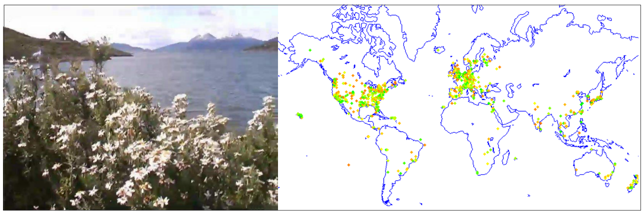

Figure 2: Probability map based on visual features

3. EXPERIMENTS & RESULTS

We tested our approach on 5091 Flickr videos from the MediaEval 2010 dataset. Five fusion experiments were evaluated as seen in table 1 by calculating the distance between the predicted point to the ground truth-point. Compared with the other approaches mentioned above, querying collaborative systems combined with the use of uploader information has the highest impact for predicting the location. Relying on either document indexing methods or visual features alone would lead to low location precision, but in combination they predict the location more precisely. The worst case is media data without location-specific information in their description, but our approach handles that problem by using low-level textual and visual similarity. As depicted in figure 3, our best fusion (guvswv) achieves an accuracy of 52.4% for city level ($\varnothing = 25km$).

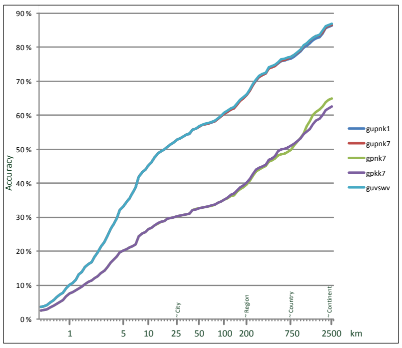

Figure 3: Accuracy against margin of error

4. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's FP7 under grant agreement number 216444 (NoE PetaMedia).

5. REFERENCES

[1] J. Hays and A. Efros. IM2GPS: estimating geographic information from a single image. pages 1–8, 2008.

[2] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 2001.

[3] B. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley LTD, 2002.

[4] P. Serdyukov, V. Murdock, and R. van Zwol. Placing flickr photos on a map. pages 484–491, 2009.