Automatic Geo-referencing of Flickr Videos

Pascal Kelm, Sebastian Schmiedeke, Kai Clüver, Thomas Sikora

Technische Universität, Berlin, Germany

E-mail: {kelm, schmiedeke, cluever, sikora}@nue.tu-berlin.de

Abstract: **We present a hierarchical, multi-modal approach for geo-referencing Flickr videos. Our approach makes use of external resources to identify toponyms in the metadata and of visual features to identify similar content. We use a database of more than 3.6 million Flickr images to group them into geographical areas and to build a hierarchical model. First, the geographical boundaries extraction method identifies the country and its dimension. Then, a visual method is used to classify the videos' location into plausible areas. Next, the visually nearest neighbour method is used to find correspondences with the training images within the pre-classified areas. As the processed video sequences are represented using low-level feature vectors from multiple key frames, we also present techniques for video to image matchings. The Flickr videos are tagged with the geo-information of the visually most similar training item within the areas previously filtered in the pre-classification step. The results show that we are able to tag one third of our videos correctly within an error margin of 1 km.**

Keywords: geo-localization, gazetteers, MPEG-7 visual features

1 INTRODUCTION

Navigating through huge databases of multimedia on the Web, especially searching for related entries, demands the use of metadata. For images or video sequences, not only textual metadata are required−geographical metadata, preferably geo-coordinates, have turned out essential for the purpose. While geo-tagging shared content has become a popular activity for users in multimedia communities, and increasing numbers of cameras automatically assign geo-coordinates to image and video data, a large majority of the resources on the Web still lack geo-tags. Consequently, automatic methods for assigning geo-coordinates to video sequences hold a large promise for improving access to video data in online communities. The main contribution of this work is a framework for geo-tag prediction that exploits both textual and visual metadata of related video data. It is shown that while visual features alone do not correlate well with locations, they successfully combine with a tag-based approach. In combination with a toponym look-up method that preselects videos by area, even low-level features of visual data improve the geo-tagging performance. The paper is structured as follows. In the next section, we cover the related work. We introduce our approach using different modalities in section 3. Results are shown in section 4, 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. [10] explain how existing standards can be combined to set up a gazetteer infrastructure allowing for bottom-up contributions 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 semantics-based information retrieval. Smart et al. [16] present a framework to access and integrate distributed gazetteer resources to build a meta-gazetteer that combines different aspects of place name data from multiple gazetteer sources. At the end they employ several similarity metrics to identify equivalent toponyms.

The approach of Hays et al. [8] is purely data-driven; their data is limited to a sub-set of Flickr images having only geographic tags. They find visually nearest neighbours to a single image based on low-level visual image descriptors and propagate the geo-location of the GPS-tagged neighbours. This approach serves as a very general means for exploring similarities between images; by itself, it provides very limiting accuracy. Working with object retrieval methods, several authors [15] [5] build visual vocabularies which are usually created by clustering the descriptor vectors of local visual features such as SIFT.

Crandall et al. [7] 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 a number of 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 key-word tags.

3 GEO-REFERENCING FRAMEWORK

Our proposed framework assigns geo-tags for Flickr videos based on their textual metadata and visual content. The idea of our method is an extension of the basic approach by Hays et al. [8] toward more powerful visual descriptions and fusion with textual metadata. In addition, we provide solutions for geo-referencing videos rather than only images.

The system includes several methods that are combined in a hierarchical way as depicted in figure 1:

Corresponding author: Pascal Kelm, TU Berlin, Einsteinufer 17, 10587 Berlin, kelm@nue.tu-berlin.de

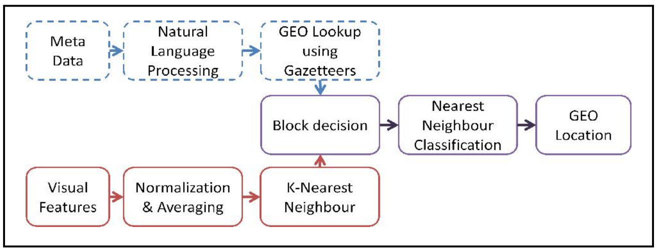
Meta Data
Natural Language Processing
GEO Lookup using Gazetteers
Block decision
Nearest Neighbour Classification
GEO Location
Visual Features
Normalization & Averaging
K-Nearest Neighbour

Figure 1: Framework overview

The first step is the pre-classification of the videos into possible areas on the world. This step is independently performed by two modules—geographical boundary extraction and a visual region model—using different modalities that are combined into a single area decision. These pre-classification modules are evaluated in different configurations in section 4. The classified areas restrict the subsequent visual similarity search of the second step to training items located in a specific area. The second module (depicted as purple boxes) uses the visual content described by visual descriptions to further predict the location. This visually nearest-neighbour method calculates the similarities between visual low-level features to assign the geo-tag of the most similar training item.

3.1 Geographical Boundaries Extraction

This approach extracts the geographical boundaries for each video sequence using the extracted toponyms of the metadata. In this approach, all promising toponyms are extracted from the user-contributed metadata of the video and then used for looking up the geo-coordinates. First, we extract the textual labelling from the video (i. e. description, title, and keywords) to collect all information about the possible location.

Then, in order to handle non-English metadata, the language is detected and the sentences are translated into English. 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 [6], 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 of possible countries in which the place designated by a particular toponym candidate may be located. The determination of a country is less ambiguous than that of a place or a city.

The geographical boundaries are determined by querying the Google Maps API [4] for the borders of the top ranked country.

3.2 Visual Region Model

For every video sequence, this method returns the visually most similar areas, which are represented by a mean feature vector of all training images and videos of the respective area.

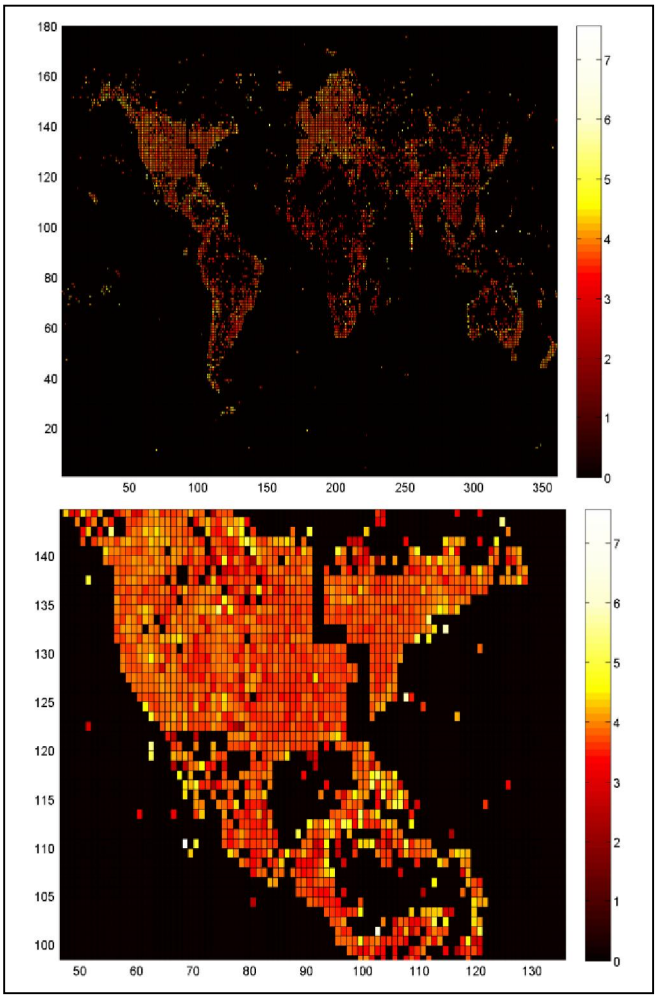

Figure 2: Visual confidence scores of a test video sequence located in Bounds Crossing (USA/Florida) placed on a map

The basic idea of the method is similar to the one described in section 3.4, but uses a mean feature vector instead of the feature vector of a single media item. An evaluation of this method is shown in section 4. For an example video[1] the visual confidence scores are shown in figure 2. Since this video sequence is captured underwater, there are many likely regions in the world based on visual features—this diving video sequence may have been recorded at any coast region in the world, and only a restriction based on textual descriptions could reduce the number of possible candidates.

3.3 Fusion for Area Decision

The methods for pre-classifying the area described in the previous sections are combined for a more accurate area classification, which also reduces the computing time in the subsequent classification step. The fusion is done in the following way:

The geographical boundary extraction (sec. 3.1) reduces the number of possible areas by restricting them to those located within the boundaries of the country detected. The Visual Region Model (sec. 3.2) returns the similarities of the concatenated feature vectors of the area model and the test video. The Euclidean norm is used for comparison of feature vectors. The area with the smallest Euclidean difference is chosen and is further analysed on video level (see next section).

3.4 Visually Nearest Neighbour

This method assigns the geo-tags of the visually most similar image within the boundaries determined by the area decision methods to the video sequence. This has the advantage that only a small subset of the training corpus needs to be processed. The method determines the visually nearest neighbour of each test video sequence within the training corpus. Since we want to reduce the temporal dimensionality of the video sequence, we use the associated key frames provided by the MediaEval placing task data set [11]. These key frames have been extracted every four seconds and their visual content is described by the following descriptors [13] using the open source library LIRE [12] with the default parameter settings: Color and Edge Directivity (CED), Gabor, Fuzzy Color and Texture Histogram (FCTH), Scalable Color (SC), Tamura, and Color Layout (CL). With these descriptors, a wide spectrum of descriptions of colour and texture within images is covered. The visual features used here are only a selection of the descriptors provided by the MediaEval set, because some of those address similar image features. The feature vectors of each descriptor are concatenated to a single feature vector for subsequent visual comparison between key frames of different videos. Since different dimensionalities and co-domains of the various descriptors render the comparison difficult, the feature vectors of each descriptor are first normalised to zero mean and unit variance.

The resulting 604-dimensional feature vector is compared to the feature vectors of the other key frames using the Euclidean norm. Other L norms were tested as well, but did not achieve better results than the L2 norm used for comparison. Since a video sequence has more than one key frame, we investigate two strategies for video-to-image comparison:

In the *keyframe-to-image* approach the video is tagged with the geo-information of the training image that has the smallest Euclidean distance to any key frame of the test video.

The *video-to-image* approach tags the video with the geo-information of the training image that contains the smallest mean Euclidean distance to all key frames of the test video.

The results of these two approaches are very similar. In the following, only the results of the video-to-image approach is shown, which performs slightly better.

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 set [11], which contains training data of about 3.6 million images and 5108 videos. The test set comprises 5108 videos.

We first discuss the impact of the geographical boundary block decision method (sec. 3.1), followed by the results of our fusion compared to the two baseline methods.

4.1 First Baseline Method (Randomness)

The first baseline method 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 item. This baseline method achieves an accuracy of about 12% for an error of 1000 km.

4.2 Second Baseline Method (Tag-Based)

The second baseline method returns the textual nearest neighbour of each video sequence using probabilistic latent semantic analysis (PLSA) on keyword tags. For this case, we choose a state-of-the-art document indexing method which applies PLSA [17], as the prediction of geographical coordinates can be regarded as retrieval of similar documents to queried tags. This unsupervised document classification method introduces a statistical latent class model to perform probabilistic mixture decomposition.

Here, the corpus of training videos is represented by a co-occurrence matrix with entries $n(w, d)$ listing the tag w in document d. The latent topic variable z associates the occurrence of tag w to document d; formally, PLSA models the probability of each co-occurrence (w, d) of words and documents as a mixture of conditionally independent multinomial distributions. See Kelm et al. [9] for further information. For the purpose of saving memory the training set is clustered and the model learning step of PLSA is applied to each cluster. The learning step is described in detail in Hofmann [17].

For each video from the training set the probability vector $P(z|d_i)$ of the respective model is calculated. In the prediction step the number of common tags of each test video to each cluster−or rather PLSA model−is determined. Then, for each

[1] Video on http://www.flickr.com/photos/62285085@N00/3484324495 located in Florida (USA)

test video the probability vector *P(z|d_{test})* of the PLSA model having most tags in common is computed:

$$P(z_l|d_{test}) = \frac{\sum_{j=1}^{M} n(d_i, w_j) P(z\_l|d_i, w_i)}{n(d_i)}$$

This PLSA test step is performed with the expectationmaximization algorithm with locked *P(w|z)* [17]. The probability vectors *P(z|d_{test})* are compared with the corresponding one in the respective cluster by applying the Euclidean distance. The test video is then labelled with the geo-tag of the most similar training video. This baseline archives an accuracy of 71% for an error of 1000 km.

4.3 Results

We investigate the impact of our hierarchical approach on the prediction performance. The hierarchical approach includes our geographical boundaries extraction method that is used to reduce the number of possible areas by querying gazetteers for extracted toponyms.

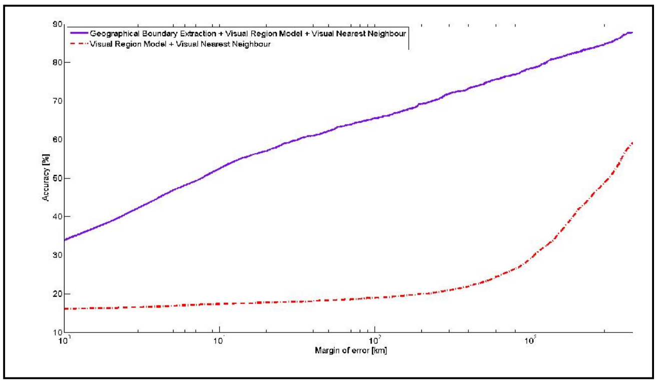

Figure 3: Accuracy plot against geographical margin of error: Usage of gazetters

The evaluation of the the pre-classification methods in combination with the visually nearest neighbour method, which selects the most similar training item within the possible areas, is shown in figure 3. The restriction made by the geographical boundaries extraction significantly increases the prediction accuracy of the visual method.

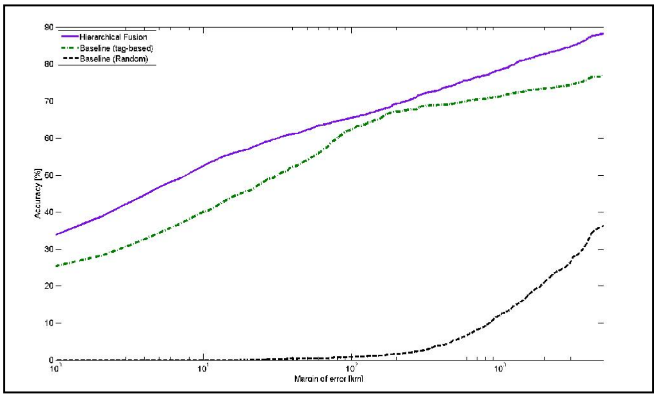

Figure 4: Accuracy plot against geographical margin of error: Our framework against baseline methods

The gain using geographical boundaries extraction amounts to up to 40% against the purely vision-based method for an error of 100 km, because the fusion of the pre-classification methods leaves only the most plausible areas.

For an example video the confidence scores for these plausible areas are shown in figure 2. Based on these confidence scores the video could be assigned to many probable areas. This geographical ambiguity is eliminated by restricting the selection to certain geographical boundaries (e.g. detected country). The accuracy is further increased by eliminating irrelevant areas.

Figure 4 shows our hierarchical approach against the two baseline methods. Our approach outperforms these baselines and achieves a considerable accuracy of 50% for an error of 8 km. This is a gain of 12% against the tag-based baseline method.

5 SUMMARY AND CONCLUSIONS

In this paper we presented a hierarchical approach for the automatic prediction of geo-tags as an improvement to previous work [9]. We presented a technique using visual and textual modalities to assign Flickr videos on the map. The fusion of textual and visual methods is important to eliminate geographical ambiguities. The external resources used—GeoNames and Wikipedia—are databases with still growing knowledge, therefore a training step is not needed. The information we use includes tags, descriptions, and titles, which can help predicting the location more precisely than using tags alone. We would like to point out that we are able to find a geo-location that is correctly located within a radius of 8 km for half of the test set.

Our proposed approach is useful for browsing and organising media items. A possible application could be automatic geo-tag suggestion in online shared media databases. Even a coarse geo-location provides the user with useful cues for finding specific landmarks.

We will improve our framework by using more distinctive visual descriptors 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.

6 ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's FP7 under grant agreement number 216444 (NoE PetaMedia). We would also like to thank the MediaEval organisers for providing the data set.

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 IEEE 12th International Conference on Computer Vision, pages 72-79, IEEE 2009.

[6] J. Baldridge. The OpenNLP Project.
http://www.opennlp.com, 2005.

[7] 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.

[8] 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, 2008.

[9] P. Kelm, S. Schmiedeke, and T. Sikora. Multi-modal, Multi-resource Methods for Placing Flickr Videos on the Map.
ACM International Conference on Multimedia Retrieval (ICMR 2011), 2011.

[10] 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.

[11] 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. ACM International Conference on Multimedia Retrieval (ICMR 2011), 2011.

[12] 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. http://www.semanticmetadata.net/lire, ACM, 2008.

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

[14] C. Manning, P. Raghavan, and H. Schütze. An Introduction to Information Retrieval. Cambridge University Press; 1 edition (July 7, 2008), 2008.

[15] 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.

[16] 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, pages 234-248, Springer 2010.

[17] T. Hofmann, Unsupervised Learning by Probabilistic Latent Semantic Analysis, in Machine Learning, vol 42, pages 177-196, Springer 2001.