How Spatial Segmentation improves the Multimodal Geo-Tagging

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

In this paper we present a hierarchical, multi-modal approach in combination with different granularity levels for the Placing Task at the MediaEval benchmark 2012. 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 bounderies detection recognizes the country and its dimension to speed up the estimation and to eliminate geographical ambiguity. Next, we prepared a training database to group them together into geographical regions and to build a hierarchical model. The fusion of visual and textual methods for different granularities is used to classify the videos' location into possible regions. At the end 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.

General Terms

placing task 2012, automatic geotagging, hierarchical segmentation

1. INTRODUCTION

The key contribution of this work is a framework for geo-tag prediction designed to exploit the relative advantages of textual and visual modalities. We will show that visual features alone show low correlation with locations but in combination with a hierarchical spatial segmentation that preselects videos into possible areas it improves the geo-tagging performance. For a detailed explanation of the Placing Task 2012 and the submitted runs, we refer to overview working notes [4].

2. FRAMEWORK

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 preclassification 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 boundaries detection extracts toponyms and uses gazetteers to increase the effectiveness of our proposed approach [3]. 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.

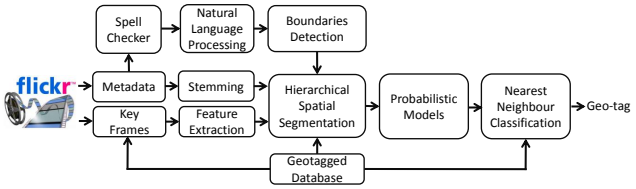

Figure 1: Textual and visual features are used in a hierarchical framework to predict the most likely location.

2.1 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.

The *boundaries extraction method* extracts the geographical 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 using Google Translate [1]. The translated metadata of the video to be geo-tagged is analysed by natural language processingin order to extract nouns and noun phrases. Each item is coarsely filtered using GeoNames [2] and Wikipedia.With the help of GeoNames, we create a rank sum of each of the possible countries and regions in which the place designated by all toponym

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

candidates could be located.

2.1.1 Textual Approach

The decision for spatial locations based on metadata can be regarded as classification of documents. For applying a probabilistic classifier we treat the spatial locations l as classes with associated metadata from the training set assigned to the spatial locations. The vocabulary V of the spatial locations includes stemmed words[1] from the tags, the titles and descriptions. 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:

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

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

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:

$$P\left(t|l\right) = \frac{N_{t,l} + 1}{\sum_{t' \in V} \left(N_{t',l} + 1\right)}, \quad (2)$$

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 the term frequency-inverse document frequency (TF-IDF). The $N_{t_k,d}$ in Eq. 1 and $N_{t,l}$ in Eq. 2 are replaced by the tf-idf scores

$$tf - idf_i = N_{t,l} \cdot \log \frac{N}{n_i}, \quad (3)$$

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

2.1.2 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.

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

2.2 Experimental Results

Our approach is focused on data fusion in a hierarchical manner, so most of our submissions considered combined results.

run1 combines textual and visual features: translation of tags and extracted words (NLP) from the title and the description. Next, Porter stemmer and stop-word elimination for each segment and granularity in the spatial segmentation. Visual Search for the k-nearest segments in the lowest hierarchy.

run2 is similar to *run1*: For the highest hierarchy level the boundaries extraction using gazetteers (GeoNames, Wikipedia) for the spell checked words is added.

run4 is using visual features for diverent granularities.

The evaluation results for different margins of error are shown in table 1.

Table 1: Results for the test set.

margin of error	run1	run2	run4
1 km	13.7 %	18.1 %	0.1 %
10 km	32.7 %	37.9 %	0.1 %
100 km	41.8 %	49.1 %	0.2 %
1,000 km	62.2 %	68 %	14.8 %
2,000 km	76.5 %	79.9 %	44.5 %
10,000 km	99.4 %	99.5 %	98.7 %
15,000 km	100 %	100 %	100 %

3. ACKNOWLEDGMENTS

We would like to acknowledge the 2011 Placing Task of the MediaEval Multimedia Benchmark for providing the data used in this research.

4. REFERENCES

[1] http://translate.google.com.
[2] http://www.geonames.org.
[3] 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.
[4] A. Rae and P. Kelm. Working notes for the placing task at mediaeval 2012. In *MediaEval Multimedia Evaluation Workshop 2012*, Santa Croce in Fossabanda Piazza Santa Croce, 5 - 56125 - Pisa - Toscana - Italia, 2012.