A Hierarchical, Multi-modal Approach for Placing Videos on the Map using Millions of Flickr Photographs

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

We present a hierarchical, multi-modal approach for placing Flickr videos on the map. Our approach makes use of external resources to identify toponyms in the metadata and of visual and textual features to identify similar content. First, the geographical boundaries extraction method identifies the country and its dimension. We use a database of more than 3.6 million Flickr images to group them together into geographical regions and to build a hierarchical model. A fusion of visual and textual methods is used to classify the videos' location into possible regions. Next, the visually nearest neighbour method uses a nearest neighbour approach to find correspondences with the training images within the pre-classified regions. The video sequences are represented using low-level feature vectors from multiple key frames. The Flickr videos are tagged with the geo-information of the visually most similar training item within the regions that is previously filtered by the pre-classification step for each test video. The results show that we are able to tag one third of our videos correctly within an error of 1 km.

Categories and Subject Descriptors

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

General Terms

geo-localization, gazetteers, Bernoulli classification, MPEG-7 visual features

1. INTRODUCTION

Geo-coordinates are a form of metadata, which is 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. 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 over the world, 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 3. The 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 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 semantics-based information retrieval. Smart et al. [16] present a framework to access and integrate distributed gazetteer resources to build a meta-gazetteer that 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. [8] is purely data-driven and 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 by Hays et al. serves as a very general means

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SBNMA'11, December 1, 2011, Scottsdale, Arizona, USA.
Copyright 2011 ACM 978-1-4503-0990-5/11/12 ...$10.00.

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 the, e. g., 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.

3. 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-tagging videos rather than only images. The system includes several methods that are combined in a hierarchical way as depicted in figure 1: The first step is the pre-classification of these videos into possible regions on the map. This step is independently performed by three modules—Geographical boundary extraction, textual region model, and visual region model—using different modalities that are combined into a single region decision. These modules are separately evaluated in section 4. The classified regions restrict the following similarity search of the second step to videos located in a specific area. The second module (depicted as purple box) 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.

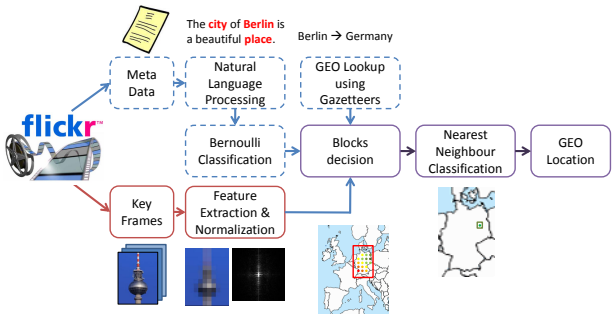


Figure 1: The hierarchical fusion of textual and visual pre-classification methods and visually nearest-neighbour classification

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 $R(c_i)$ 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. This in case, the pre-selection relies only on the following two pre-classification methods.

The geographical boundaries for a detected country are determined by querying the Google Maps API [4] for borders. The resulting geographical boundaries supports the visually nearest neighbour (sec. 3.5) search in terms of preselecting possible near-located video sequences.

3.2 Textual Region Model

The decision for geographical region based on tags can be regarded as classification of documents. For applying a text classifier we segment the world map into 360×180 regions according to the meridians and parallels. These regions are considered as classes C. The geo-tagged images with associated tags from the training set are assigned to the 360×180 classes according to geographical segmentation. The vocabulary V of the image tags from the training set is generated and leads to a Bernoulli model of tags within the regions. This model generates an indicator for each tag t_i of the vocabulary, which is either 1 indicating presence of the tag in the region or 0 indicating absence. The probabilities $p(t_i|C)$

for each tag t_i belonging to each of the regions C are calculated from the Bernoulli model. Since zeros within the $p(t_i|C)$ probabilities cause problems during probability calculation, $P(t|C)$ is smoothed by adding ones [14]:

$$P(t|C) = \frac{N_{ct} + 1}{\sum_{t' \in V} (N_{ct'} + 1)},$$

where N_{ct} is the number of training images containing the tag t in the class c (resp. region).

For classifying the test video sequences d_j into regions C, their tags are used in a naive bayes classifier excluding prior knowledge of region probabilities $P(C)$:

$$\log P(d_j|C) = \sum_i \log p(t_i|C).$$

The use of logarithmic probabilities solves the problem of arithmetic underflow by applying addition instead of multiplication of probabilities in the Bayes's rule [14]. The region with the highest logarithmic probability $\log P(d_j|C)$ is chosen for further analysis. The following two images [figure 2] shows for a video[1] the textual confidence scores.

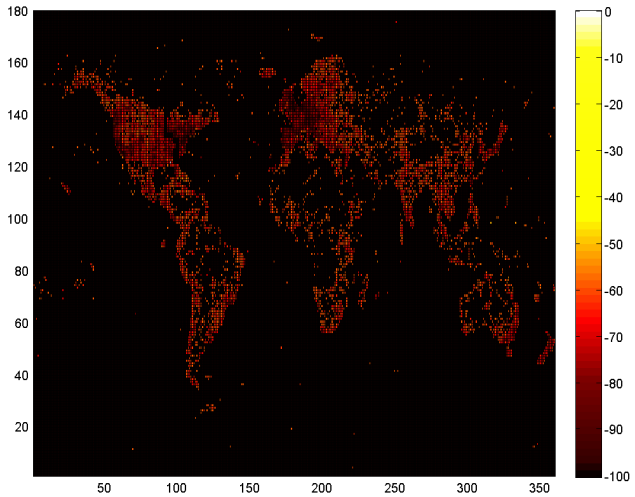
3.3 Visual Region Model

For every video sequence, this method returns the visually most similar regions, which are represented by a mean feature vector of all training images and and videos of the respective region. The basic idea of the method is similar to the one described in section 3.5, but uses a mean feature vector instead of the feature vector of a feature vector of a single media item. An evaluation of this method is shown in section 4. This model was also evaluated using a median feature vector instead of a mean feature vector. As it turned out, using the median feature vector did not increase the prediction results. For the example video the visual confidence scores are shown in figure 3. 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.

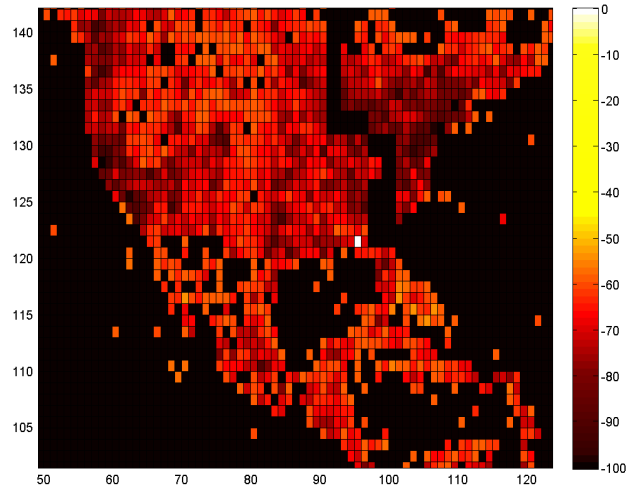
3.4 Fusion for Region Decision

The methods for pre-classifying the region described in previous sections are combined for a more accurate region 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 regions by restricting them to those located within the boundaries of the detected country, if sufficient metadata was available. The Textual Region Model (sec. 3.2) returns the log-likelihoods of the remaining regions to the tags of each test video. The Visual Region Model (sec. 3.3) returns the similarities of the concatenated feature vectors of the region model and the test video. The Euclidean norm is used for comparison of feature vectors. The results of both models are combined into single score. The combination of visual and textual region models is difficult, since the confidence scores are different in scale and range. Thus the fusion is done on rank level by using the rank sum algorithm. The region with the highest rank is chosen and is further analysed on video level (see sec. 3.5).

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



(a) Textual confidence scores

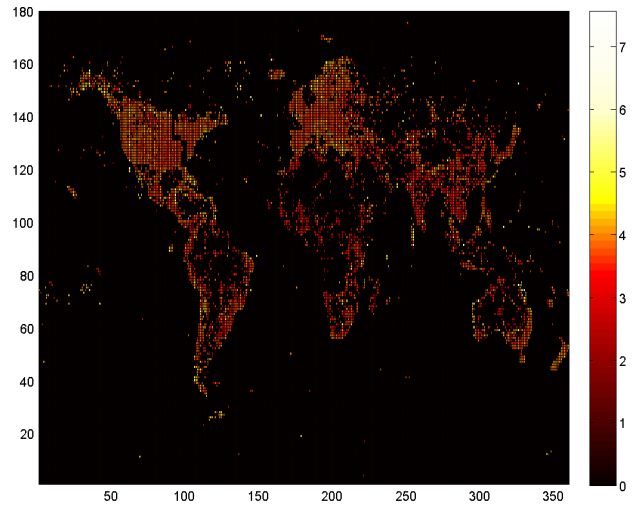


(b) Detail of (a)

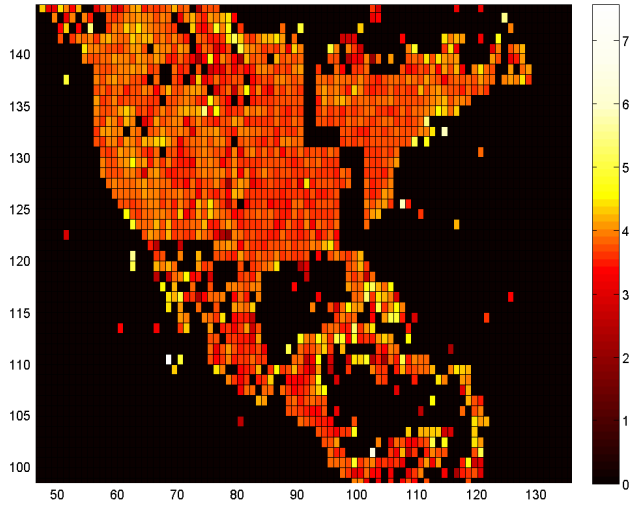
Figure 2: Textual confidence scores of a test video sequence located in Bounds Crossing (USA/Florida)

3.5 Visually Nearest Neighbour

This method assigns the geo-tags of the visually most similar image within the boundaries determined by the region decision methods to the video sequence. This has the advantage that only a small subset of the training corpus needs to be computed. 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 [11] placing task data set. 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. These visual features are only a selection of the descriptors provided by the MediaEval set,



(a) Visual confidence scores



(b) Detail of (c)

Figure 3: Visual confidence scores of a test video sequence located in Bounds Crossing (USA/Florida)

because some of them 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 previously normalised to zero mean and unit variance. The resulting feature vector has a dimension of 604 and is compared to the feature vectors of the other key frames using the Euclidean norm. Other L norms did not achieve better results than the L^2 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 tags the video 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 tag 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. So 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] containing training data of about 3.6 million images and 5108 videos. This dataset contains all information about the user, the geolocation, the video and all textual information. We first discuss the results of the each single block decision method (sec. 3.1-3.3), followed by the results of the fusion against the baseline.

Our results are compared to a baseline method that is based on randomness to show the statistical significance. For this purpuse, 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.

We investigate the variation of the prediction performance of our hierarchical approach with changes in the pre-classification step:

- Geographical boundaries extraction is used to reduce the number of possible regions by querying gazetteers for extracted toponyms.
- Textual region models are applied to choose the regions with the highest probabilities.
- The Visual region models is used as pre-classification step to choose the visually most similar regions.

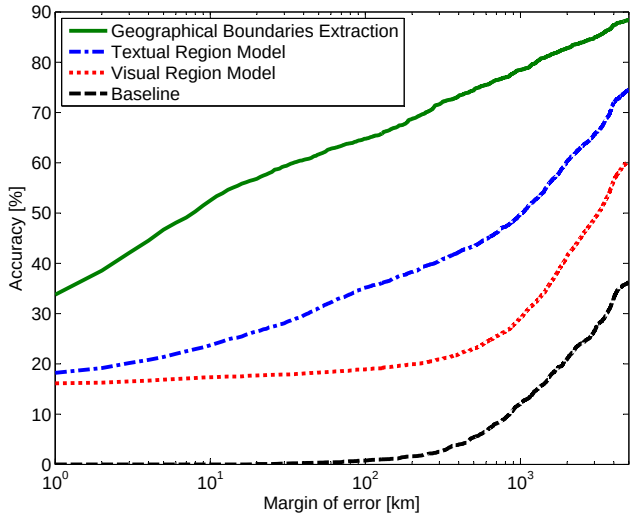


Figure 4: Accuracy plot against geographical margin of error: Pre-classification methods in combination with visually nearest neighbour

For evaluation we used the orthodromic distance which is the shortest distance between any two points on the surface

of a sphere measured along a path on the surface of the sphere. The single evaluation of the these pre-classification methods in combination with the visually nearest neighbour method, which selects the most similar training item within the possible regions is shown in figure 4 and in table 1.

Table 1: Results of each pre-classification methods in combination with visually nearest neighbour on selected margin of errors

Exp.	5 km	25 km	50 km	100 km	250 km
GBE	46.69%	58.18%	61.66%	64.78%	70.12%
TRM	21.37%	27.38%	31,06%	35.16%	39.60%
VRM	16.85%	17.81%	18.26%	18.89%	20.37%

The gain using geographical boundaries extraction amounts to up to 40% against the purely vision-based method for an error of 100 km. When user tags are not available in the flickr video this geographical boundaries are open for the whole dataset. The fusion of the pre-classification methods leaves only the most probable regions within the geographical boundaries to be selected. For a video located in Florida (USA) the textual and visual confidence scores are shown in figure 2 and 3. Based on these confidence scores the video could be assigned to many probable regions.

This geographical ambiguity is eliminated by superimposing these scores and restricting the selection to certain geographical boundaries (e.g. detected country). These few selected regions reduce the computation complexity during the similarity calculation. In addition, the accuracy is further increased by eliminating irrelevant regions.

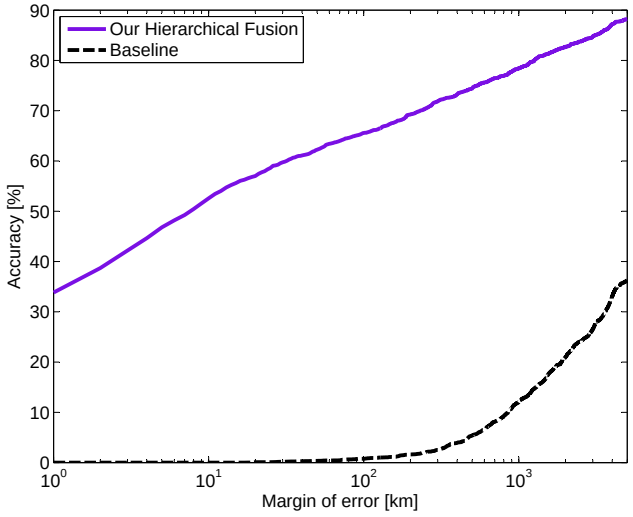


Figure 5: Accuracy plot against geographical margin of error: Fusion

Figure 5 shows our hierarchical approach against the baseline method. The hierarchical approach uses a fusion of mentioned pre-classification methods and the visually nearest neighbour method to predict geo-coordinates. In the fusion step the confidence scores of the textual region model and the visual region model are superimposed with the extracted geographical boundaries to determinate the most likely geographical region. Our approach achieves an accuracy of 50% for an error of 8 km.

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] , where a fallback system was used when direct amiguilty elimination failed. In this paper, visual and textual modalities are used to determinated likely regions and post-classify with the aid of visual descriptors within these regions. We presented a two stage technique to assign Flickr videos on the map using visual and textual modalities. The geographical boundary extraction as pre-classifier is a useful method for eliminating irrelevant regions. This method is based on tags, descriptions, and titles, that contains more pieces of information than using tags alone. The reduction to few region candidates also reduced the computational time during the similarity calculation within the regions. The worst case for placing media data is the lack of location-specific information in their metadata, but our approach handles that problem by using low-level textual and visual similarity. At the end it is shown that the fusion of textual and visual methods is important to eliminate geographical ambiguities. The results suggest that our proposed approach would be quite useful for browsing and organising media items. 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. In our future work, we would use object recognition algorithms that can be applied to media items to predict locations almost accurate to the metre, i.e. a photograph depicting the Eiffel Tower can be tagged precisely using gazetteers (like images of the geo-located 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) and 261743 (NoE VideoSense). We would also like to thank the MediaEval organisers for providing this data set.

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] 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] 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.

[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. ACM, http://www.semanticmetadata.net/lire/, 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.