Multimodal Location Estimation on Flickr Videos

Gerald Friedland, Jaeyoung Choi, Howard Lei, Adam Janin
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704-1198
{fractor, jaeyoung, hlei, janin}@icsi.berkeley.edu

ABSTRACT

The following article describes an approach to determine the geo-coordinates of the recording place of Flickr videos based on both textual metadata and visual cues. The system is tested on the MediaEval 2010 Placing Task evaluation data, which consists of 5091 unfiltered test videos. The system presented in this article is less complex, uses less training data, and is at the same time more accurate than the best system presented in the evaluation in August 2010. The performance peaks at being able to classify 14 % of the videos with less than 10 m accuracy. The article describes the realization of the system, analyses of the different uses of multimodal cues and gazetteer information.

Categories and Subject Descriptors

H3.1 [**Information Storage and Retrieval**]: Indexing methods; I4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Sensor Fusion*

General Terms

Experimentation

Keywords

Video, Tagging, Multimodal, Location Estimation, Content Analysis

1. INTRODUCTION

Today's computers have begun to have the computational power and memory to be able to process large amounts of data in different modalities, such as audio, video, and text. This, in combination with the large amount of multimedia data freely accessible in the Internet, provides an opportunity to improve the robustness of current multimedia content analysis approaches and attack problems that were impossible to solve before.

A multimedia content analysis task that has only recently become tractable to research is the estimation of the location of a video recording that lacks geo-location metadata. The task is sometimes called "multimodal location estimation" or "placing". Just as a human analyst uses multiple sources of information and context to determine geo-location, it is obvious that for location estimation, the investigation of clues across different modalities and the combination with diverse knowledge sources from the web can lead to better results than investigating only one stream of sensor input (e.g. reducing the task to an image retrieval problem).

The task has recently caught the attention of researchers in the multimedia, signal processing, and machine learning communities because of the large amount of available geo-tagged media on the Internet that could be used as training data, allowing algorithms to work on data volumes rarely seen before. In addition, the task is hard enough to require the collaboration of many different experts and communities, which is a challenge on its own.

This article describes an approach to determine the geo-coordinates of the recording place of Flickr videos based on textual metadata and visual cues. The system is tested on the MediaEval 2010 Placing Task evaluation data and proves to be less complex, while using less training data, and being more accurate than the best system presented in the evaluation in August 2010. The performance peaks at being able to classify 14 % of the videos with less than 10 m accuracy.

2. DEFINITION AND MOTIVATION

As initially defined in [6], *multimodal location estimation* denotes the utilization of one or more cues potentially derivable from different media, e.g. audio, video, and textual metadata, to estimate the geo-coordinates of content recorded in digital media.

Note that the location of the shown content might not be identical to the location where the content was created, in fact in most cases there is a bias because the camera records GPS coordinates of the location where the camera is located not of the objects captured. For practical purposes, the research presented in here focusses on finding one unique location per file, even if the video happens to be edited to show different locations. In such a case, the location shown in the median frame of the video is estimated.

Work in the field of location estimation is currently creating progress in many areas of multimedia research. As discussed in [6], cues used to estimate location can be extracted using methods derived from current research areas. Since found data from the Internet is used, multimodal location estimation work is performed using much larger test and training sets than traditional multimedia content analysis

MM'11, November 28–December 1, 2011, Scottsdale, AZ

tasks and the data is more diverse as the recording sources and locations differ greatly. This offers the chance to create machine learning algorithms of potentially higher generality. Overall, multimodal location estimation has the potential to advance many fields, some of which we don't even know of as they will be created based on user demand for new applications. However, apart from the academic motivation there are several real-world incentives behind the attempt to solving multimodal location estimation.

Location-based services are rapidly gaining traction in the online world. Besides major players like Google and Yahoo!, there are many smaller start-ups in the space as well. The main driving force behind these services is the enabling of a very personalized experience. In a parallel development, a growing number of sites now provide public APIs for structured access to their content, and many of these already come with geo-location functionality. Flickr, YouTube, and Twitter all allow queries for results originating at a certain location. Likewise, the believe is that retro-fitting archives with location information will be attractive to many businesses and enables new usage scenarios. Also, except for specialized solutions, GPS is not available indoors or where there is no line of sight with satellites. So multimodal location estimation helps provide geo-location where it is not regularly available. For example, vacation videos and photos could now be grouped even if GPS data is not attached. As discussed in [13], this was one of the main motivations in the MediaEval evaluation. Movie producers have long searched for methods to find scenes at specific locations or showing specific events in order to be able to reuse them.

3. RELATED WORK

Given the motivation to solve this task described in the previous section, it is no wonder that initial approaches to location estimation have already started several years ago. In earlier articles [18, 22], the location estimation task is reduced to a retrieval problem on self-produced, location-tagged image databases. The idea is that if the image is the same then the location must be the same too. In other work [8], the goal is to estimate just a rough location of an image taken as opposed to close-to-exact GPS location. For example, many pictures of certain types of landscapes can occur only on certain places on Earth. Krotkov's approach [3] for robot applications, extracts sun altitudes from images while Jacobs' system [9] relies on matching images with satellite data. In both of these settings single images have been used or images have been acquired from stationary webcams. In the work of [12], the geo-location is also determined based on the estimate of the position of the sun. They provide a model of photometric effects of the sun on the scene, which does not require the sun to be visible in the image. The assumption, however, is that the camera is stationary and hence only the changes due to illumination are modeled. This information in combination with time stamps is sufficient for the recovery of the geo-location of the sequence. A similar path is taken in [10].

Previous work that has been carried out in the area of automatic geotagging of multimedia that has based on tags have also been mostly carried out on Flickr images. User-contributed tags have a strong location component, as brought out by [20], who reported that over 13 % of Flickr image tags could be classified as locations using Wordnet. In [17], the geo-locations associated with specific Flickr tags are predicted using spatial distributions of tag use. A tag which is strongly concentrated in a specific location has a semantic relationship with that location. User-contributed tags are exploited for geotagging by [19], who use tag distributions associated with locations represented as grid cells on a map of the Earth is used to infer the geographic locations of Flickr images. The approach in [7] and [1] reports on combining visual content with user tags. However, the accuracy in [7] is only reported with a minimum granularity of 200 km.

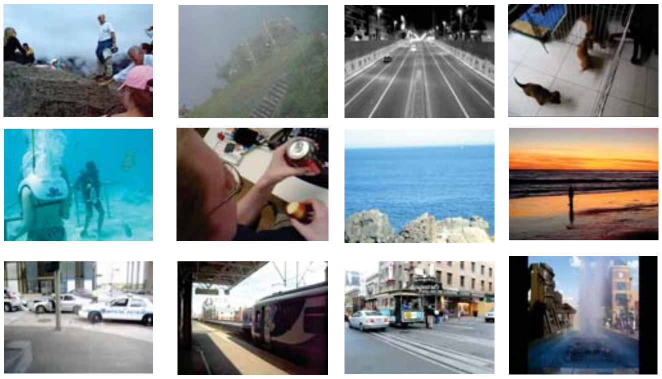

Figure 1: Several frames from the MediaEval 2010 test set as described in Section 4.

Multimodal location estimation on videos has been first defined and attempted in [6] where the authors match ambulance videos from different cities, even whithout using textual tags. The first evaluation on multimodal location estimation on randomly selected consumer-produced videos has been performed in the 2010 MediaEval Placing task [13]. Several notable systems participated in the evaluation [21, 11, 2, 4, 16], including the predecessor of the system described herein. The rules of the evaluation prohibit us to compare and rank the system results as of the evaluation. Please refer to the cited references for further information. However, the system presented in this article is less complex, uses less training data, and is at the same time more accurate than the best system presented in the evaluation in August 2010

4. DATASETS

4.1 MediaEval 2010 Dataset

The MediaEval 2010 Placing Task data set consists of Creative Common-licensed videos that were manually crawled from Flickr. The videos are in MPEG-4 format and include the Flickr metadata in XML format. The metadata for each video includes user-contributed title, tags, description, comments and also information about the user who uploaded the videos. Additionally, the metadata also includes information about the user's contacts, favorites, and all videos uploaded in the past. The data set was divided into training data (5091 videos) and test data (5125 videos).

According to [13], videos were selected both to provide a broad coverage of users, and also because they were geotagged with a high accuracy at the "street level". Accuracy shows the zoom level the user used when placing the photo on the map. There are 16 zoom levels, and these correspond to 16 accuracy levels (e.g., "region level", "city level", "street

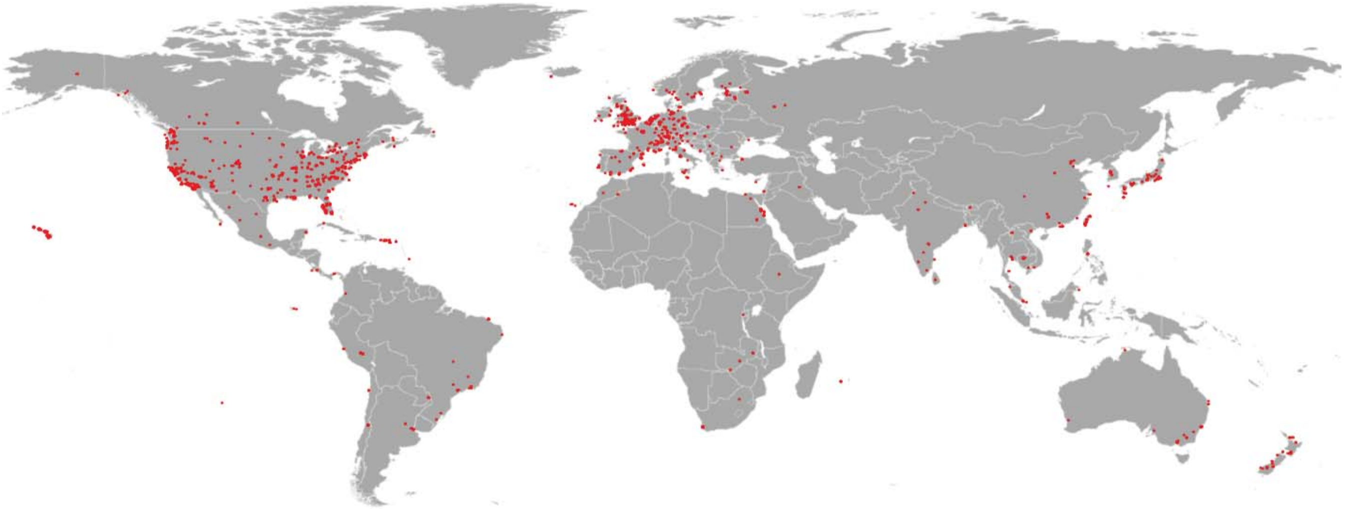

Figure 2: Distribution of the videos of the MediaEval 2010 Placing Task development set. As discussed in Section 4, randomly sampling videos from Flickr results in a non-uniform geographical prior.

level"). The sets of users from the test and the training collections were disjoint in order to not introduce a user-specific bias. This bias will be discussed further in Section 6. In order to allow visual matching as performed in [8], the dataset also contained metadata and features extracted from 3,185,258 Flickr images. However, not all the photos had textual metadata and the photos were only guaranteed to have geo-tagging at least region level accuracy.

4.2 Characteristics of the Data

Flickr requires that an uploaded video must be created by its uploader (if a user violates this policy, Flickr sends a warning and removes the video). Manual inspection of the data set lead us initially to conclude that most of visual/audio contents lack reasonable evidence to estimate the location without textual metadata. For example, many videos were recorded indoors or in a private space such as a backyard of a house. This indicates that the videos are not pre-filtered or pre-selected in any way to make the data set more relevant to the task, and are therefore likely representative of videos selected at random. Figure 1 shows several sample frames from the MedieEval 2010 test set.

However, metadata provided by the user often provides direct and sensible clues for the task. 98.8 % of videos in the training set were annotated by their uploaders with at least one title, tags, or description, often including location information.

4.3 Additional Data

Because of the non-uniformity of the MediaEval 2010 training and test set, we used additional data to make the training data more equally distributed over the earth. In addition to the MediaEval 2010 data, we also included the data used for the experiments described in [8], which was legal as of the rules of the benchmark. The data originally consists of 6.4 million images from Flickr categorized into countries and states (in case of US). We sampled pictures from each region and used their unique Flickr photo ID to download the metadata from Flickr. 759,249 metadata records were collected in this way. Furthermore, we collected additional photos from Flickr by dividing the area of the earth into 1 km grid cells, counting the number of photos for each grid cell. If the cell contained more than 15 photos, we sampled 15 % of photos. This resulted in about 1,131,698 new metadata records and photos. All metadata was collected and saved in the same format as the MediaEval photo dataset UserID, PhotoID, HTML link to photo, latitude and longitude, tags, date taken, and date uploaded. Again, we ensured that the user set stays disjoint between training and test set.

5. TECHNICAL APPROACH

Our approach is a data-driven multimodal method that uses both the textual tags as well as visual features. The input is a test video with metadata. From the metadata, we only use the user-annotated tags (not the title, or descriptions) that are included in the metadata record for each Flickr video or photo. We also experimented with using title and descriptions but the results were significantly worse than only using the tags. Furthermore, 2601 of the 5125 videos in the test data did not contain a description. The algorithm is described below.

First we process the tags. For each given tag in the test video record, we determine the spatial variance by searching the training data for an exact match of the tag and creating a list of the geo-locations of the matches. If only one location is found, the spatial variance is trivially small. We pick the centroid location of the top-3 tags with the smallest spatial variance. This results in 0 to 3 coordinates. In the case of 0 coordinates (e.g. because the video is not tagged or no tags match), we assume the most likely geo-coordinate based on the prior distribution of the MediaEval traing set (see Figure 2), which is the point with lattitude and longitude (40.71257011, -74.10224), a place close to New York City.

For the visual processing step, the input is the median frame of the test video and the 1 to 3 coordinates of the previous step. We resize the frame to 128 × 128 pixels and extract gist [15] features and color histogram. The gist descriptor is based on a 5 × 5 spatial resolution with each bin containing responses to 6 orientation and 4 scales. The color histograms were created based on the CIELAB transformed

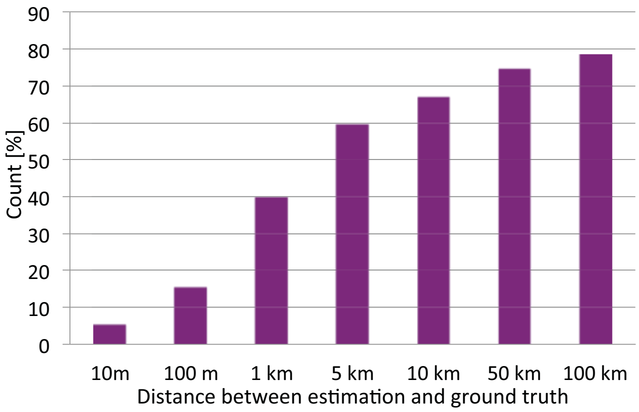

Figure 3: The resulting accuracy of the algorithm as described in Section 5.

pixels for the frame, like in [8]. The histogram has 4 bins for L, and 14 bins for A and B. We then adopt the matching methodology from [8]. We used Euclidean distance to compare gist descriptors and chi-square distance for color histograms. Weighted linear combination of distances was used as the final distance between frames. The scaling of the weights was learned by using a small sample of the training set and normalizing the individual distance distributions so that each the standard deviation of each of them would be similar. We use 1-NN matching between the test frame and the all the images in a 100 km radius around the 1 to 3 coordinates from the previous step. We pick the match with the smallest distance and output its coordinates as a final result.

This multimodal algorithm is less complex than previous algorithms (see Section 3), yet produces more accurate results on less training data. The following section analyses the accuracy of the algorithm and discusses experiments to support individual design decisions.

6. RESULTS AND ANALYSIS

The evaluation of our results is performed by applying the same rules and using the same metric as in the MediaEval 2010 evaluation. In MediaEval 2010, participants were to built systems to automatically guess the location of the video, i.e., assign geo-coordinates (latitude and longitude) to videos using one or more of: video metadata (tags, titles), visual content, audio content, and social information. Even though training data was provided (see Section 4), any "use of open resources, such as gazetteers, or geo-tagged articles in Wikipedia was encouraged" [14]. The goal of the task was to come as close as possible to the geo-coordinates of the videos as provided by users or their GPS devices. The systems were evaluated by calculating the geographical distance from the actual geo-location of the video (assigned by a Flickr user, creator of the video) to the predicted geo-location (assigned by the system). While it was important to minimize the distances over all test videos, runs were compared by finding how many videos were placed within a threshold distance of 1 km, 5 km, 10 km, 50 km and 100 km. For analyzing the algorithm in greater detail, here we also show distances of below 100 m and below 10 m. The lowest distance category is about the accuracy of a typical GPS localization system in a camera or smartphone.

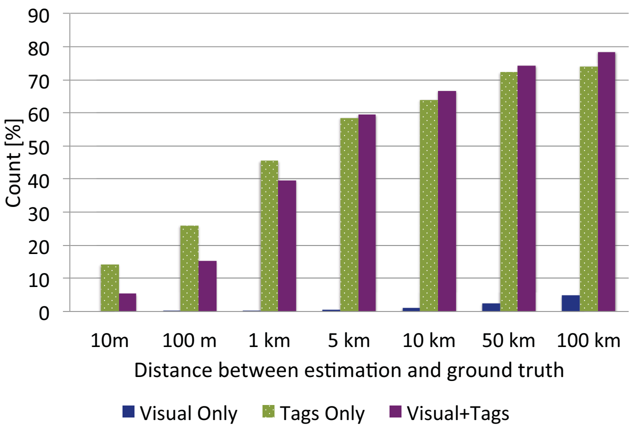

Figure 4: The resulting accuracy when comparing tags-only, visual-only, and multimodal location estimation as discussed in Section 6.1.

First we discuss the results as generated by the algorithm described in Section 5. The results are visualized in Figure 3. The results shown are superior in accuracy than any system presented in MedieEval 2010. Also, although we added additional data to the MediaEval training set, which was legal as of the rules explained above, we added less data than other systems in the evaluation, e.g. [21]. Compared to any other system, including our own, the system presented here is the least complex.

6.1 About the Visual Modality

Probably one of the most obvious questions is the impact of the visual modality. As a comparison, the image-matching based location estimation algorithm in [8] started reporting accuracy at the granularity of 200 km. As can be seen in Figure 4, this is consistent with our results: Using the location of the 1-best nearest neighbor in the entire database compared to the test frame results in a minimum accuracy of 10 km. In contrast to that, tag-based localization reaches accuracies of below 10 m. For the tags-only localization we modified the algorithm from Section 5 to output only the 1-best geo-coordinates centroid of the matching tag with lowest spatial variance and skip the visual matching step. While the tags-only variant of the algorithm performs already well, using visual matching on top of the algorithm decreases the accuracy in the finer-granularity ranges but increases overall accuracy, as in total more videos can be classified below 100 km. Out of the 5091 test videos, using only tags 3774 videos can be estimated correctly with an accuracy better 100 km. The multimodal approach estimates 3989 correctly in the range below 100 km.

6.2 The Influence of Non-Disjoint User Sets

Each individual person has his own idiosyncratic method of choosing a keyword for certain events and locations when they upload videos to Flickr. Furthermore, the spatial variance of the videos uploaded by one user is low on average. At the same time, a certain amount of users uploads many videos. Therefore taking into account to which user a video

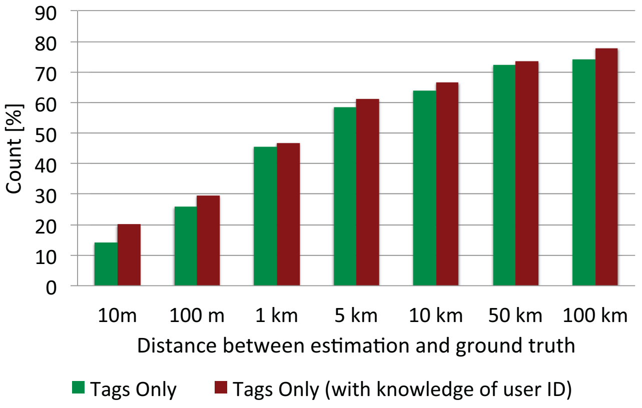

Figure 5: The resulting accuracy when taking into account user locality as discussed in Section 6.2.

belongs seems to have a higher chance of finding geographically related videos. For this reason, videos in the MediaEval 2010 test dataset were chosen to have a disjoint set of users from the training dataset. However, the additional training images provided for MediaEval 2010 are not user disjoint with the test videos. Therefore we are able to run an experiment exploiting the user overlap. Instead of searching for a matching keyword in all videos within the dataset, we limit the search to just the videos uploaded by the same user, cutting down on confusion. If the user is not in the training image set, we use the tags-only algorithm as described in the previous paragraph. The results are shown in Figure 5. As can be seen, the accuracy is increased significantly, especially in the regions below 1 km. While exploiting user locality is legal as of the rules of MediaEval 2010, it is generally considered bad practice. The Flickr dataset often contains many videos and photos by the same individual and exploiting this property of the database might not be helpful to solve the multimodal location estimation problem in general.

6.3 Using a Geographical Gazetteer

As discussed in Section 3, related work has tried using geographical gazetteers to increase the robustness of the search. Also, Flickr provides the home location of the user of an uploaded video which could be treated as an equivalent to a user-based gazetteer as every user can be mapped to a place on earth. We therefore performed experiments to see if the incorporation of this type of semantic information would be useful. We used the open service *Geonames.org*. GeoNames covers all countries and contains 8 million entries of placenames and corresponding geo-coordinates. It provides a web-based search engine and an API which returns a list of matching entries ordered by their relevance to the query.

After the filtering procedures to pick toponym candidates from the textual metadata, we passed the query to the Geonames search engine and retrieved the list of possible matches. We added the entity with the highest relevance (the first entity in the list) to the list of candidate entities.

Choosing the best match among the obtained list of candidate locations was similar to the method we used in Section 5. We plot all candidate entities on a map and pick the one that has the largest count of neighbors with lowest spatial variance. For a detailed description of the procedure including the filtering and the tie-breaking process, see [2].

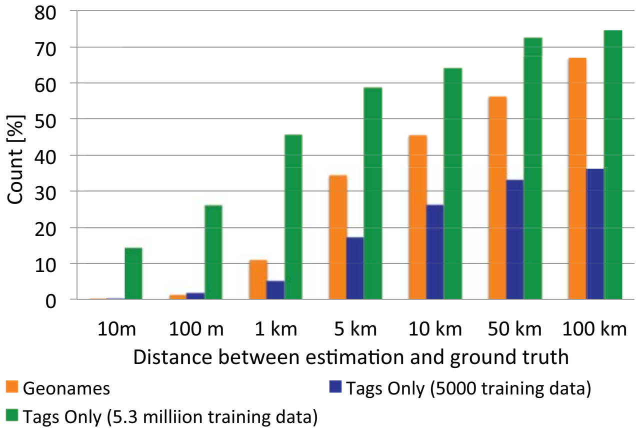

Figure 6: Comparing the use of a geographical gazetteer versus the technical approach in Section 5 with different training data volumess. See also discussion in Section 6.3.

We found that incorporating gazetteer information can help significantly with sparse datasets. However, with enough sample records, tag matching as described in Section 5 outperforms the gazetteer approach, even when incorporating the Flickr-specific home location as described above. Figure 6 shows the results comparing tag matching and using Geonames plus a user's home location.

7. FUTURE DIRECTIONS

When the user provides sufficient tags in the metadata and the tags are location-specific to where the video was taken, our approach shows potential to return the location very accurately. In fact, our algorithm already outperforms the availabilty of explicitly geo-tagged multimedia (e.g. as EXIF data), as only about 5 % of Flickr videos and images are geo-tagged [5] and our tags-only approach is already able to classify about 14 % of the videos within the 10 m range.

Of course, there are some videos which confusingly contain toponyms in their metadata to describe an incident or an object which is not proximal to where the video was recorded (e.g. ”Goodbye Oregon, hello San Francisco”). While not an exception, these cases are much more difficult. We expect that further integration with other media will help here. As shown in [6] both visual and acoustic information contained in the videos can contribute to the task. MediaEval 2011 will also include a tasks that does not allow the usage of textual metadata.

Given the simplicity of the algorithm, reducing complexity is not the first priority. However, an important goal should be to reduce data usage of the algorithm further while maintaining or increasing accuracy. This would especially help making the algorithm more efficient since a full run of the textual/visual system is about 75 hours on a single CPU.

Another major problem is uniformity of the training data. In order to create a generalized system, it should be based on samples for every country and region. However, as discussed earlier, for various reasons, several countries and regions are not represented by geo-tagged videos or images at all. A first step here would be to take non-geotagged images and videos and use Amazon's Mechanical Turk to annotate their locations manually so that they can be used for training. In addition, the development of unsupervized training approaches and confidence metrics would help to tackle this problem further.

8. CONCLUSION

In this article we described a system for the estimation of the recording location of Flickr videos. The system uses both tags as well as video content and achieves significant accuracy improvements due to the integration of the two media. The accuracy of the system is higher than any of the systems presented in the MediaEval evaluation in August 2010. At the same time, our approach relies on less data and its realization seems to be the least complex compared to related work. The article also discusses several experiments that contribute to the understanding and validity of the task. Even though visual information alone does not seem to be competitive compared to a tags-only approach, the combination of the two media does improve the overall performance. We verified the claim that overlapping users in test and training sets will introduce a bias which might hinder the generalizability of approaches. Finally, we discussed the use of gazetteer data and found that semantic technologies like that can be helpful but mostly in situations were not enough training data is available. Finally, we pointed to future directions. We believe future research will mostly focus on improving the accuracy of audio/visual approaches, which can be investigated both with and without the use of additional textual tags.

Futher information about the project can be found at http://mmle.icsi.berkeley.edu.

Acknowledgments

This research is supported by an NGA NURI grant #HM11582-10-1-0008.

9. REFERENCES

[1] L. Cao, J. Yu, J. Luo, and T. S. Huang. Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression. In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pages 125–134, New York, NY, USA, 2009. ACM.

[2] J. Choi, A. Janin, and G. Friedland. The 2010 ICSI Video Location Estimation System. In *Proceedings of MediaEval*, October 2010.

[3] F. Cozman and E. Krotkov. Robot localization using a computer vision sextant. In *IEEE international conference on robotics and automation*, pages 106–106, 1995.

[4] D. Ferres and H. Rodriguez. TALP at MediaEval 2010 Placing Task: Geographical Focus Detection of Flickr Textual Annotations. In *Proceedings of MediaEval*, October 2010.

[5] G. Friedland and R. Sommer. Cybercasing the Joint: On the Privacy Implications of Geo-Tagging. In *Proc. USENIX Workshop on Hot Topics in Security*, August 2010.

[6] G. Friedland, O. Vinyals, and T. Darrell. Multimodal Location Estimation. In *Proceedings of ACM Multimedia*, pages 1245–1251, 2010.

[7] A. Gallagher, D. Joshi, J. Yu, and J. Luo. Geo-location inference from image content and user tags. In *Proceedings of IEEE CVPR*. IEEE, 2009.

[8] J. Hays and 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, 2008.

[9] N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless. Geolocating static cameras. In *IEEE international conference on computer vision*, pages 1–6, 2007.

[10] I. Junejo and H. Foroosh. Estimating geo-temporal location of stationary cameras using shadow trajectories. *Computer Vision–ECCV 2008*, pages 318–331, 2008.

[11] P. Kelm, S. Schmiedeke, and T. Sikora. Video2GPS: Geotagging using collaborative systems, textual and visual features: MediaEval 2010 Placing Task. In *Proceedings of MediaEval*, October 2010.

[12] J. Lalonde, S. Narasimhan, and A. Efros. What does the sky tell us about the camera? *Computer Vision–ECCV 2008*, pages 354–367, 2008.

[13] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and G. J. Jones. Automatic Tagging and Geo-Tagging in Video Collections and Communities. In *ACM International Conference on Multimedia Retrieval (ICMR 2011)*, page to appear, April 2011.

[14] Mediaeval web site. *http://www.multimediaeval.org*.

[15] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[16] J. Perea-Ortega, M. Garcia-Cumbreras, L. Urena-Lopez, and M. Garcia-Vega. SINAI at Placing Task of MediaEval 2010. In *Proceedings of MediaEval*, October 2010.

[17] T. Rattenbury and M. Naaman. Methods for extracting place semantics from Flickr tags. *ACM Transactions on the Web (TWEB)*, 3(1), 2009.

[18] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*, pages 1–7, 2007.

[19] P. Serdyukov, V. Murdock, and R. van Zwol. Placing Flickr photos on a map. In *ACM SIGIR*, pages 484–491, 2009.

[20] B. Sigurbjoernsson and R. V. Zwol. Flickr Tag Recommendation based on Collective Knowledge. In *ACM WWW*, pages 327–336, April 2008.

[21] O. Van Laere, S. Schockaert, and B. Dhoedt. Ghent University at the 2010 Placing Task. In *Proceedings of MediaEval*, October 2010.

[22] W. Zhang and J. Kosecka. Image based localization in urban environments. In *3D Data Processing, Visualization, and Transmission, Third International Symposium on*, pages 33–40, 2006.