The 2011 ICSI Video Location Estimation System

Jaeyoung Choi
International Computer Science Institute
1947 Center St., Suite 600
Berkeley, CA 94704, USA
jaeyoung@icsi.berkeley.edu

Howard Lei
International Computer Science Institute
1947 Center St., Suite 600
Berkeley, CA 94704, USA
hlei@icsi.berkeley.edu

Gerald Friedland
International Computer Science Institute
1947 Center St., Suite 600
Berkeley, CA 94704, USA
fractor@icsi.berkeley.edu

ABSTRACT

In this paper, we describe the International Computer Science Institute's (ICSI's) multimodal video location estimation system presented at the MediaEval 2011 Placing Task. We describe how textual, visual, and audio cues were integrated into a multimodal location estimation system.

1. INTRODUCTION

The MediaEval 2011 Placing Task [6] is to automatically estimate the location (latitude and longitude) of each query video using any or all of metadata, visual/audio content, and/or social information. For a detailed explanation of the task, please refer to the Placing Task overview paper [6]. Please note that the videos for the Placing Task were not filtered or selected for content in any way and represent "found data". This is described in more detail in [1] and [4]. The system presented herein utilizes the visual and acoustic content of a video together with textual metadata, whereas the system from 2010 [1] only leveraged metadata. As a result, the accuracy has improved significantly compared to 2010. The system is described as follows.

2. SYSTEM DESCRIPTION

Our system integrates textual metadata with visual and audio cues from the video content into a multimodal system. Each component and the overall integration is described as follows.

2.1 Utilizing textual metadata

From all available textual metadata, we only utilized the user-annotated tags and ignored the title and descriptions.

Our intuition for using tags to find the geolocation of a video is the following: If the spatial distribution of a tag based on the anchors in the development data set is concentrated in a very small area, the tag is likely a toponym (location name). If the spatial variance of the distribution is high, the tag is likely something else but a toponym. For a detailed description of our algorithm, see [1]. Also, we use GeoNames [7], a geographical gazetteer, in permitted runs as a backup method when the spatial variance algorithm returns 0 coordinate.

2.2 Utilizing visual cues

In order to utilize the visual content of the video for location estimation, we reduce location estimation to an image retrieval problem, assuming that similar images mean similar locations. We therefore extract GIST features [5] for both query and reference videos and run a k-nearest neighbor search on the reference data set to find the video frame or a photo that has is most similar. GIST features have been shown to be effective in automatic geolocation of images [3]. We convert each image and video frame to grayscale and resize them to 128×128 pixels before we extract a GIST descriptor with a 5×5 pixels spatial resolution with each bin containing responses to 6 orientation and 4 scales. We use Euclidean distance to compare the GIST descriptors and use 1-nearest neighbor matching between the closest pre-extracted frame to the temporal mid-point of a query video and all photos and frames from the reference videos.

2.3 Utilizing acoustic cues

Our approach for utilizing acoustic features is based on [4]. The article showed the feasibility and super-human accuracy of acoustic features for location estimation by describing a city identification system derived from a state-of-the-art 128-mixture GMM-UBM speaker recognition system, with simplified factor analysis and Mel-Frequency Cepstral Coefficient (MFCC) features. For each audio track, a set of MFCC features is extracted and one Gaussian Mixture Model (GMM) is trained for each city, using MFCC features from all its audio tracks (i.e. city-dependent audio tracks). This is done via MAP adaptation from a universal background GMM. The log-likelihood ratio of MFCC features from the audio track of each query video is computed using the pre-trained GMM models of each city. A likelihood score of each query video corresponding to each of the cities is obtained. A city with the highest score is picked as the query video's location. This approach, however, limits the range of estimated locations to pre-picked 15 cities around the world with the highest concentration of videos. This was due to the relatively small amount of 10,000 videos provided compared to more than 3 millions images and metadata.

2.4 Multimodal integration

Although recent research on automatic geolocation of images using visual and acoustic features have shown to be promising (e.g., [3, 2, 4]), the performance of these experiments is not in the same ballpark as the ones using textual metadata. When cues from multiple modalities are used together, we found that textual metadata provided by the user plays a dominant role in providing cues for the placing task. Therefore, our system is designed to use visual

MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy

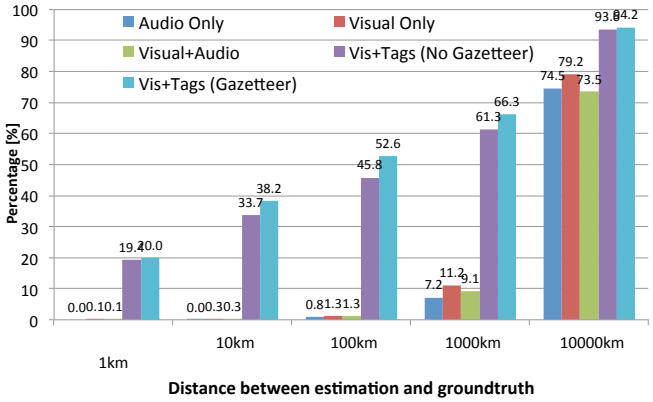

Figure 1: Comparison of runs result as described in Section 3.

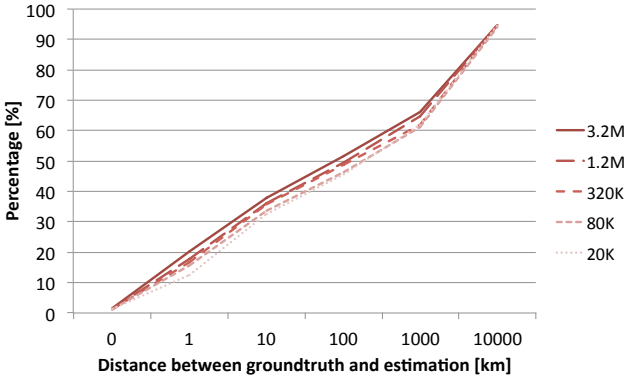

Figure 2: Increasing the size of development data improves performance

features as second preference and acoustic features only as third preference.

In order to integrate the visual features, we first run the tags-only algorithm from Section 2.1 and use the resulting top-3 tags as anchor points for a 1-NN search using visual features (see Section 2.2). We compare against all reference images and video frames within 1 km radius from the 3 anchor points. As explained above (see Section 2.3), the number of audio references is much smaller and the result of the audio matching is always one of 15 pre-defined cities. Therefore, the acoustic approach was only used as a backup when the visual distance between query video and any reference video was too large (i.e., the algorithm was unable to find a similar enough scene).

3. RESULTS AND DISCUSSION

Figure 1 shows the comparative result of our runs using audio only, visual feature only, audio+visual feature and tag+visual feature approaches. As explained above, the tag-based approach shows far better performance than other approaches that does not use textual metadata. Also, given the amount of reference data, the gazetteer information does

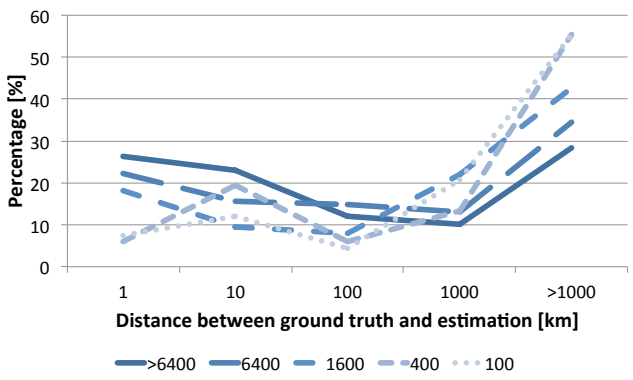

Figure 3: Test videos from denser region has higher chance of being estimated within closer range from groundtruth.

seem to contribute to the accuracy, although very little. With so little data available for audio matching, acoustic cues did not seem to contribute to the performance significantly when used alone or together with visual feature as described in Section 2.4.

Figure 2 shows that using more development data helps, especially boosting the number of correct estimation within 1km radius of ground truth. A little over 14% of the test data don't contain any useful information at all in the metadata (tag, title, and description). The training curve of test videos that were left over after applying the text based algorithm (not shown here due to the lack of space) shows the curve reaching 14.6% when 3.2 million development data were used.

Figure 3 shows that the system works better in dense areas compared to sparse areas. The whole map was divided into approximately 100 km by 100 km grid and the number of development data was counted for each grid.

In conclusion, we believe that the biggest challenge for the future is being able to handle sparse reference data.

4. REFERENCES

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

[2] 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.

[3] 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.

[4] H. Lei, J. Choi, and G. Friedland. City-Identification on Flickr Videos Using Acoustic Features. In *ICSI Technical Report TR-11-001*, April 2011.

[5] 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.

[6] A. Rae, V. Murdock, and P. Serdyukov. Working Notes for the Placing Task at MediaEval 2011. In *MediaEval 2011 Workshop*, Pisa, Italy, September 2011.

[7] M. Wick. Geonames. *http://www.geonames.org*, 2011.