Preliminary Exploration of the Use of Geographical Information for Content-based Geo-tagging of Social Video

Xinchao Li, Claudia Hauff, Martha Larson, Alan Hanjalic
Delft University of Technology
Delft, The Netherlands
{x.li-3,c.hauff,m.a.larson,a.hanjalic}@tudelft.nl

ABSTRACT

Estimating the geo-location of an image or video is an interesting and challenging task in information retrieval and computer vision. In this paper, a pure image content based approach for this task is described. We partition the world map into regions based on external data sources (climate and biomes data). We hypothesize that such a partition yields regions of high visual stability, which in turn is hypothesized to improve geolocation estimation. The exploratory experiments that we designed and carried out on the MediaEval Placing Task 2012 test data were not able to confirm these hypotheses.

1. INTRODUCTION

With the rapid development of GPS-equipped cameras, increasing numbers of photos and videos are labelled with their geo-locations, however, most existing photos/videos are not geo-labelled. The research question of the Placing Task is how to estimate the geo-location of a video[1], given its image attributes, audio information and all available metadata. The task is described in more detail in [5].

2. SYSTEM DESCRIPTION

The proposed system focuses on how to derive location information from the visual content of videos. As can be seen from previous work [1, 4], estimating the location based on the visual content alone is a difficult task. However, it is worthwhile investigating, as not all videos contain user-provided tags (35.7% of test videos of the 2012 data set do not contain any tags while 13.1% contain a single tag only). Moreover, we envision the visual-based estimate to be combined with the text-based estimate for videos where metadata text is available.

We aim to partition the world map into regions (for classifier training purposes) in such a way that there is a high visual stability within a region and a high visual variability between regions. As an example consider the two regions of the Great Victoria Desert and the South Pole—the images that are taken within one region will be highly similar (in a way that we believe can be detected by visual features), while images that are taken in the two regions are likely to be visually very distinct. To find suitable region boundaries, we rely on two external data sources: climate data and biomes data. Both provide alternative partitions of the world map.

In order to explore the visual content, we use the key frames of videos, and represent each frame by its visual features. These features provide a global description of the images, and were extracted by the placing task organisers using the open source library LIRE [5]. To estimate each test video's location, the world is divided into several subregions. We assume that each region has a certain visual stability and that they can be distinguished from each other. To represent each region, 3 million photos in the training set are assigned to their respective region according to their geo-location. Then, one model is trained for each region to assign test videos to the most probable region. This formulations constitutes a multi-class classification problem. A support vector machine with RBF kernel is utilized for this task.

3. RUN DESCRIPTION

According to the different division methods, we implemented three runs:

RUN 1: As a baseline, the world map is divided similarly to [3] based on the density distribution of training photos. First, the entire world map is represented by a single region. Training images are added iteratively. Once photos in one region exceed a threshold, the region is split into four subregions. This splitting procedure stops once a region reaches the lower size limit.

RUN 2: Climate information is utilized to divide the world with respect to different temperature regions. The global climate data consists of temperature data from 7278 temperature stations around the world. We use the annual average temperatures to divide the world into 6 temperature regions: from -20°C to 40°C with 10°C intervals. Temperature stations in each temperature region are then clustered into subsets according to their geo-locations. Training photos are then distributed to the corresponding subset by assigning them to the closest temperature station.

RUN 3: Anthropogenic biomes, which describe the terrestrial biosphere in its human-altered form using global ecosystem units [2], are explored to divide the world

[1] For our purposes, we consider a video to consist of a sequence of images (keyframes).

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

map. These biomes were identified and mapped using a multi-stage procedure based on population, land use and land cover. As these biomes are sparsely distributed around the world, training photos are first assigned to biomes, and then photos within a biome are clustered into subregions.

4. RESULTS

In our experiments, test videos are first assigned to one region, and a separate prediction for the individual key frames of each test video is generated. Since frames in one video may yield different predictions, a soft voting strategy is used to make the final prediction for the whole video:

$$\tilde{r} = \mathop{\arg\max}_{r \in Regions} \sum_{\forall i} P_r(i) \qquad (1)$$

where $P_r(i)$ is the predicted probability of the ith frame belonging to region r.

Within one region, key frames of the test video are matched among all training photos in that region, and the best matched training photo's geo-location is propagated to the test video. The results are evaluated by calculating the distance between predicted and actual geo-locations, and by determining the number of videos whose locations were estimated to be within a given distance threshold of the actual location.

Table 1: Run results (4182 videos): number of test videos located within $\{1, 10, 100, 1000, 5000\}$km of the ground truth.

	<1km	<10km	<100km	<1000km	<5000km
Run1	0	0	8	186	925
Run2	0	0	5	112	746
Run3	2	3	21	375	1458

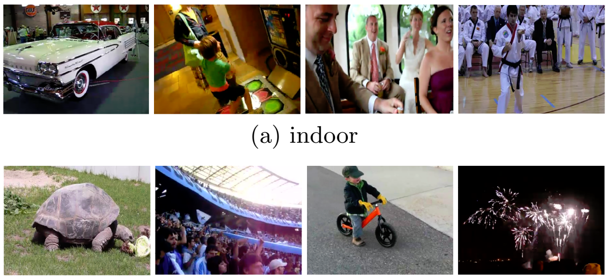
The run results are presented in Table 1. In general, our method to explore visual information does not result in a reliable location estimator. To reveal the visual stability within one region, we further analysed the region accuracy, i.e., the number of videos that were assigned to the correct *region*. As shown in Table 2, compared to a random prediction, all three runs achieve similar low region accuracy. As subregions in the same region type (e.g., sharing same biome) may look similar, we further conduct another experiment which only use photos to train and test, and use original biomes in Run3, but do not split each biome into subregions. This formulations constitutes a 22-class classification problem. The classification accuracy is 12.17%, which is higher than random guess, 04.55%. This indicates that the biome regions own some weak visual stability.

Table 2: Region accuracy: percentage of test videos assigned to the correct region.

	Run1	Run2	Run3
SVM	0.62%	0.10%	0.14%
Random Prediction	0.45%	0.21%	0.15%

5. DISCUSSION

We further manually checked the content of the test videos by randomly selecting the middle key frames of 500 videos from the 4182 videos (12%). As shown in Figure 1, about 216 videos (42%) are indoor videos, which contain very little visual location related information. Within the outdoor videos, half of them focus on an object or event (e.g., animal, football game or fireworks) and thus also contain few visual clues to estimate the location. We conjecture that this fact has a huge impact on the proposed system. As the features used to represent the video content are global features, they are not suitable for finding similar local objects.



(a) indoor

(b) outdoor

Figure 1: Sample of 2012 test video frames

To determine the significance of this issue, we randomly sampled 458 training photos. We found 126 (27.5%) of them to be indoor photos. Recall that our initial assumption was as follows: we can divide the world map into regions that have a high within-region visual stability and a high between-region variability. In this case, indoor images in the training set are noisy information that are likely to mislead the training of SVM classifiers.

In conclusion, our proposed methods for dividing the world map in combination with the utilized visual features are not able to provide a reasonable geo-location prediction for the given test videos. Future work will focus on (i) how to exploit the qualitative insights gained (i.e., the indoor/outdoor distribution), (ii) how to exploit the user-provided metadata, and, (iii) how to combine the visual and textual features in a principled manner.

6. REFERENCES

[1] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 761–770, New York, NY, USA, 2009. ACM.

[2] E. C. Ellis and N. Ramankutty. Putting people in the map: anthropogenic biomes of the world. In *Frontiers in Ecology and the Environment*, 6(8), pages 439–447, 2008.

[3] C. Hauff and G.-J. Houben. WISTUD at MediaEval 2011: Placing Task. In *MediaEval'11*, pages 1–1, 2011.

[4] P. Kelm, S. Schmiedeke, and T. Sikora. Multi-modal, multi-resource methods for placing Flickr videos on the map. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 52:1–52:8, New York, NY, USA, 2011. ACM.

[5] A. Rae and P. Kelm. Working notes for the Placing Task at MediaEval 2012. In *MediaEval 2012 Workshop*, Pisa, Italy, October 4-5 2012.