The 2014 ICSI/TU Delft Location Estimation System

Jaeyoung Choi[1,2], Xinchao Li[2]
[1]International Computer Science Institute, Berkeley, CA, USA
[2]Multimedia Computing Group, Delft University of Technology, Netherlands
[1]jaeyoung@icsi.berkeley.edu, [2]x.li-3@tudelft.nl

ABSTRACT

In this paper, we describe the ICSI/TU Delft video location estimation system presented at the MediaEval 2014 Placing Task. We describe two text-based approaches based on spatial variance and graphical model framework, a visual-content-based geo-visual ranking approach, and a multi-modal approach that combines the text and visual-based algorithms.

1. INTRODUCTION

The Placing Task 2014 [2] is to automatically estimate the geo-location of each query video using any or all of metadata, visual/audio content, and user information. For the text-based approaches, we used the spatial variance based baseline system [3] and the graphical model based framework [1] that poses the geo-tagging problem as one of inference over the graph. The graphical model jointly estimates the geo-locations of all the test videos, which helps obtain performance improvements. The visual-based location estimation is based on the evidence collected from images that are not only geographically close to the query's location but it also exploits the visual similarity to the query image within the considered image collection [4]. For the fusion of these systems' results, we ran both systems and chose the result of the text-based system as the overall result, except when the confidence was low, in which case we chose the visual-based result.

2. SYSTEM DESCRIPTION

2.1 Text-based Approach

2.1.1 Spatial Variance approach

The intuition behind this approach is that 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. 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 [3]. This approach was used as a baseline to evaluate the performance of the graphical model based algorithm. For each query, the confidence of the estimation was represented by e^{-v^2} where v^2 is the lowest spatial variance of the keywords. From all available textual metadata, we utilized the user-annotated tags, and title. Machine tags were treated the same way as the user-annotated tags. This also applies to the following graphical model based approach.

2.1.2 Graphical model based approach

The random variables in our graphical model setup are the geo-locations of the query videos that need to be estimated [1]. We treat the textual tags as observed random variables that are probabilistically related to the geo-location of that video. The goal is to obtain the best estimate of the unobserved random variables (locations of the query videos) given all the observed variables. We used graphical models to characterize the dependencies amongst the different random variables and use efficient message-passing algorithms to obtain the desired estimates.

An undirected graphical model or a Markov Random Field (MRF) $G(V, E)$ consists of a vertex set V and an edge set E. The vertices (nodes) of the graph represent random variables $\{x_v\}_{v \in V}$ and the edges capture the conditional independencies amongst the random variables through graph separation. The joint probability distribution for a N-node pairwise MRF can be written as follows:

$$p(x_1, ..., x_N) = \prod_{i \in V} \psi(x_i) \prod_{(i,j) \in E} \psi(x_i, x_j). \quad (1)$$

$\psi(.)$'s are known as potential functions that depend on the probability distribution of the random variables.

Given the training data, we fit a Gaussian Mixture Model (GMM) for the distribution of the location given a particular tag t, i.e., $p(x|t)$. The intuition is that tags usually correspond to one or more specific locations and the distribution is multi-modal (e.g., the tag "washington" can refer to two geographic places). Given that for many of the tags, the GMM will have one strong mixture component, the distribution $\psi(x_i)$, can be approximated by a Gaussian distribution with the mean $(\tilde{\mu}_i)$ and variance $(\tilde{\sigma}_i^2)$ given by,

$$(\tilde{\mu}_i, \tilde{\sigma}_i^2) = \left(\frac{\sum_{k=1}^{n_i} \frac{1}{\sigma_i^{k2}} \mu_i^k}{\sum_{k=1}^{n_i} \frac{1}{\sigma_i^{k2}}}, \frac{1}{\sum_{k=1}^{n_i} \frac{1}{\sigma_i^{k2}}} \right), \quad (2)$$

where μ_i^k and σ_i^{k2} are the mean and variance of the mixture component with the largest weight of the distribution $p(x_i|t_i^k)$. The location estimate for the ith query video $\hat{x}_i$ is taken to be $\tilde{\mu}_i$ and the variance $\tilde{\sigma}_i^2$ provides a confidence metric on the location estimate.

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

	10m	100m	1km	10km	100km	1000km
run1	0.24	3.15	16.65	34.70	45.58	60.67
run2	0.17	1.60	3.88	5.86	6.82	17.43
run3	0.22	2.75	16.28	46.20	52.81	72.19
run4	0.31	3.41	12.13	19.95	22.82	33.79
run5	0.30	3.12	12.75	24.82	27.33	42.89
Oracle	0.41	4.52	19.05	37.02	47.86	65.81

Table 1: Percentage of correctly estimated query images/videos of each run

2.2 Visual content-based approach

For the visual-based location estimation, we propose the Geo-Visual Ranking (GVR) approach [4]. The basic intuition is that, compared to the images from the wrong location, more images from the ground truth location will likely contain more elements of the visual content of the query image. Thus, instead of choosing the nearest neighbor image, or relying on the biggest cluster of visual neighbors of the query image, we searched for geo-visual neighbors of the query image. Geo-visual neighbors are images that are sufficiently visually similar to the query image and also taken at the same location as the query image. Let's assume a case where a query image has two visually similar geo-tagged images taken at different locations (which we refer to as *candidate images*). The nearest neighbor approach faces difficulty in this situation as the probability to select the wrong reference image from the two candidates is high. However, the GVR approach's estimation is affected by additional sets of images that are found around at both *candidate images*' locations (referred to as *candidate geo-visual neighbors* at *candidate locations*). These *candidate geo-visual neighbors*' contribution to the decision is based not on just the number of the images in each neighbor, but on the combined visual proximity to the query image, aggregated over all images from a set. Use of the set's visual proximity makes it possible to point to the right candidate image even if it has a smaller set of geo-neighbors than the candidate image. We used SURF descriptors extracted using the BoofCV software with the default parameters and used exact k-means to cluster these descriptors and generate visual words.

2.3 Multimodal Approach

For the fusion of these systems' results, we ran both systems, and for the text-based estimations with low confidence, visual-based result was used instead. The optimal threshold for confidence was searched using grid search over the development set and the used value was when the variance v^2 was 25.

3. RESULTS AND DISCUSSION

We submitted four runs: *run1* - spatial variance approach (text only), *run2* - visual-content-based geo-visual ranking approach, *run3* - graphical model based approach (text only), *run4* - spatial variance + GVR, and *run5* - graphical model based approach + GVR. Each column in Table 1 shows what percentage of test videos and images were placed within 10m, 100m, 1km, 10km, 100km, and 1000km from the ground truth location. For the text-based approaches (*run1* and *run3*), graphical model approach performed slightly worse than the spatial variance approach in locating videos within 10m, 100m, and 1km from the ground truth location. It outperformed the spatial variance approach in other ranges by a large margin. One theory behind this result is that the graphical model's belief propagation process causes the query node to move away from the ground truth location as the reference images or videos that are far away from the ground truth have influences that are more than desired. For the both text-based approaches, we ignored the description of the photo/video as its usage degraded the performance.

The visual-based approach (*run2*) has lower accuracy in all ranges when compared to the text-based approaches (*run1* and *run3*). However, note that visual-only result does relatively well in the lower error range (10m, 100m, and 1km). This implies that local feature matching gives very good estimation when similar image can be found in the training set.

For the multimodal approaches (*run4* and *run5*), replacing the text-based estimation with a low confidence score with the visual-based estimation helped improving the system's performance in the 10m and 100m range. *Oracle* in Table 1 shows the result of the oracle-condition experiment where we chose the estimation between the *run1* and *run2* with the shorter error distance. It shows the upper bound of the multimodal approach and the possible margin of performance increase is high. Future work needs to investigate an optimal method for the fusion of multimodal features.

4. ACKNOWLEDGMENTS

This work was partially supported by funding provided to ICSI through National Science Foundation grant IIS:1251276 ("BIGDATA: Small: DCM: DA: Collaborative Research: SMASH—Scalable Multimedia content AnalysiS in a High-level language"). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors or originators and do not necessarily reflect the views of the National Science Foundation.

5. REFERENCES

[1] J. Choi, G. Friedland, V. Ekambaram, and K. Ramchandran. Multimodal location estimation of consumer media: Dealing with sparse training data. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pages 43–48. IEEE, 2012.

[2] J. Choi, B. Thomee, G. Friedland, L. Cao, K. Ni, D. Borth, B. Elizalde, L. Gottlieb, C. Carrano, R. Pearce, and D. Poland. The placing task: A large-scale geo-estimation challenge for social-media videos and images. In *Proceedings of the 3rd ACM International Workshop on Geotagging and Its Applications in Multimedia*, 2014.

[3] G. Friedland, J. Choi, H. Lei, and A. Janin. Multimodal Location Estimation on Flickr Videos. In *Proceedings of the 3rd SIGMM Workshop on Social Media in Conjunction with ACMMM*, 2011.

[4] X. Li, M. Riegler, M. Larson, and A. Hanjalic. Exploration of feature combination in geo-visual ranking for visual content-based location prediction. In *MediaEval*, 2013.