The 2012 ICSI/Berkeley Video Location Estimation System

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

Venkatesan Ekambaram, Kannan Ramchandran
Department of EECS, UC Berkeley
Berkeley, CA 94704, USA
{venkyne, kannanr}@eecs.berkeley.edu

ABSTRACT

In this paper, we describe the ICSI/Berkeley video location estimation system presented at the MediaEval 2012 Placing Task. We describe the graphical model based framework that poses the geo-tagging problem as one of inference over the graph. Our algorithm jointly estimate the geo-locations of all the test videos, which helps obtain performance improvements over existing algorithms in the literature that process each query video independently. Our modeling provides a generic theoreteical framework that can be used to incorporate any other available textual, visual or audio features.

1. INTRODUCTION

The Placing Task [2] is to automatically estimate the geolocation 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 [2].

2. SYSTEM DESCRIPTION

2.1 Baseline approach using textual metadata

From all available textual metadata, we utilized the user-annotated tags, and title. We ignored the descriptions as their usage degraded our performance. This also applies to the graphical model based approach. The 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. 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]. This approach was used as a baseline to evaluate the performance of the graphical model based algorithm. We used Yahoo! Maps geocoder in permitted runs to obtain the geo-location from the uploader's locality information.

2.2 Graphical model based approach

Graphical models provide an efficient representation of dependencies amongst different random variables and have been extensively studied in the statistical learning theory community [3]. The random variables in our setup are the geo-locations of the query videos that need to be estimated. We treat the textual tags and visual and audio features 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 use 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 [3]. The joint probability distribution for a N-node pairwise MRF can be written as follows [3],

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

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

We use the sum-product algorithm to find approximate maginals. In the sum-product algorithm, messages are passed between nodes that take the following form:

$$m_{j\to i}(x_i) \propto \int_{x_j} \psi(x_i, x_j)\psi(x_j) \prod_{k\in N(j)/i} m_{k\to j}(x_j)dx_j, \quad (2)$$

In order to obtain a graphical model representation for our problem setup, we need to model the joint distribution of the query video locations given the observed data. We use a simplistic conditional dependency model for the random variables as described below. Each node in our graphical model corresponds to a query video and the associated random variable is the geo-location of that query video. Intuitively, if two images are nearby, then they should be connected by an edge since their locations are highly correlated. The problem is that we do not know the geo-locations a priori. However, given that textual tags are strongly correlated to the geo-locations, a common textual tag between two images is a good indication of the proximity of geo-locations. Hence, we will build the graphical model by having an edge between two nodes if and only if the two query videos have at least one common textual tag. Note that this textual tag need not appear in the training set. The model could be further improved using the audio and visual features as well.

Let x_i be the geo-location of the ith video and $\{t_i^k\}_{k=1}^{n_i}$ be the set of n_i tags associated with this video. Based on our model, under a pairwise MRF assumption, the joint probability distribution factorizes as follows:

$$p(x_1, ..., x_N|\{t_1^k\}, ..., \{t_N^k\}) \propto \prod_{i\in V} \psi(x_i|\{t_i^k\}) \prod_{(i,j)\in E} \psi(x_i, x_j|\{t_i^k\}, \{t_j^k\}).$$

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

Given the potential functions, one could use the sum-product iterates (2), to estimate $p(x_i|\{t_1^k\}, ..., \{t_N^k\})$.

We now need to model the node and edge potential functions. 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). To estimate the parameters of the GMM, we use an algorithm based on Expectation Maximization that adaptively chooses the number of components for different tags using a likelihood criterion. Assuming conditional independence for different tags, we take the node potential as follows, $\psi(x_i) \propto \prod_{k=1}^{n_i} p(x_i|t_i^k)$. For the potential functions, $\psi(x_i, x_j|\{t_i^k\}, \{t_j^k\})$, we use a very simple model. Intuitively, if the common tag between two query videos i and j occurs too frequently either in the test set or the training set, that tag is most likely a common word like "video" or "photo" which does not really encode any information about the geographic closeness of the two videos. In this case, we assume that the edge potential is zero (drop edge (i, j)) whenever the number of occurrences of the tag is above a threshold. When the occurrence of the common tag is less frequent, then it is most likely that the geographic locations are very close to each other and we model the potential function as an indicator function,

$$\psi(x_i, x_j|\{t_i^k\}, \{t_j^k\}) = \begin{cases} 1 & \text{if } x_i = x_j, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

This model is a hard-threshold model and we can clearly use a soft-version wherein the weights on the edges for the potential functions are appropriately chosen.

Further, we propose the following simplification, which leads to analytically tractable expressions for the potential functions and message updates. 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,

$$(\tilde{\mu}_i, \tilde{\sigma}_i^2) = \left(\frac{\sum_{k=1}^{n_i} \frac{1}{\sigma_i^{k2}} \mu_i^k}{\sum_{k=1}^{n_i} \frac{1}{\sigma_i^{k2}}}, \frac{1}{\sum_{k=1}^{n_i} \frac{1}{\sigma_i^{k2}}} \right), \quad (4)$$

where μ_i^k and σ_i^{k2} are the mean and variance of the mixture component with the largest weight of the distribution $p(x_i|t_i^k)$. Under this assumption, the iterations of the sum-product algorithm take on the following simplistic form. Node i at iteration m, updates its location estimate $(\hat{\mu}_i(m))$ and variance $(\hat{\sigma}_i^2(m))$ as follows,

$$\hat{\mu}_i(m) = \frac{\frac{1}{\tilde{\sigma}_i^2}\tilde{\mu}_i + \sum_{j \in N(i)} \frac{1}{\hat{\sigma}_j^2(m-1)} \hat{\mu}_j(m-1)}{\frac{1}{\tilde{\sigma}_i^2} + \sum_{j \in N(i)} \frac{1}{\hat{\sigma}_j^2(m-1)}}, \quad (5)$$

$$\hat{\sigma}_i^2(m) = \frac{1}{\frac{1}{\tilde{\sigma}_i^2} + \sum_{j \in N(i)} \frac{1}{\hat{\sigma}_j^2(m-1)}}. \quad (6)$$

The location estimate for the ith query video $\hat{x}_i$ is taken to be $\hat{\mu}_i(m)$ at the end of m iterations, or when the algorithm has converged. The variance $\hat{\sigma}_i^2(m)$ provides a confidence metric on the location estimate.

2.3 Utilizing visual cues

We ran a K-nearst neighbor search using extracted Gist feature on the reference data set to find the video frame or a photo that is most similar. Detailed description is in ICSI's 2011 system working note [1].

3. RESULTS AND DISCUSSION

	1km	10km	100km	1000km	10000km
run1	728	996	1321	1951	3719
run2	683	1316	1964	2445	3581
run3	740	1043	1396	1970	3731
run4	0	4	27	419	3160

Table 1: run1 - baseline approach without using gazetteeer, run2 - graphical model based approach with gazetteer, run3 - baseline approach with gazetteer, run4 - k-NN with gist visual feature

Table 1 shows our results of the four submitted runs. Each column shows how many videos out of 4182 test videos were placed within 1km, 10km, 100km, 1000km, and 10000m from the ground truth location. Although graphical model approach with gazetteer (run2) performed slightly worse than baseline approach in locating videos within 1km from the true location, it outperformed baseline approaches (run1 and run3) in other range by a large margin.

In the real world, only 5 percent of Internet videos are geo-tagged, and hence the training set is much smaller than the test set, contrary to what is assumed otherwise in the literatures. Also, the training data is biased toward certain regions such as those with dense population or touristy spots. The analysis of a data-driven algorithm in different regions with varying data densities show that grids with a denser population of training data perform significantly better than those with lesser training data. Thus, an estimation model is needed to handle the sparsity. The graphical model approach handles the sparsity in the training data by intelligently processing the test videos in such a way that each additional test (query) video not only is placed but also improves the quality of the existing training data.

4. ACKNOWLEDGMENTS

This research is partly funded by NSF EAGER grant IIS-1138599, NSF Award CNS-1065240 and a KFAS Doctoral Study Abroad Fellowship.

5. REFERENCES

[1] J. Choi, H. Lei, and G. Friedland. The 2011 ICSI Video Location Estimation System. In *Proceedings of MediaEval*, Septermber 2011.

[2] A. Rae and P. Kelm. Working Notes for the Placing Task at MediaEval 2012. In *MediaEval 2012 Workshop*, Pisa, Italy, October 2012.

[3] M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1:1–305, 2008.