Late-Breaking Developments in the Field of Artificial Intelligence
Papers Presented at the Twenty-Seventh AAAI Conference on Artificial Intelligence

Localizing Web Videos from Heterogeneous Images

Xian-Ming Liu
Beckman Institute
University of Illinois at Urbana-Champaign
xliu102@illinois.edu

Yue Gao
School of Computing
National University of Singapore
kevin.gaoy@gmail.com

Rongrong Ji
Department of Cognitive Science
Xiamen University, P.R. China
rrji@xmu.edu.cn

Shiyu Chang
Beckman Institute
University of Illinois at Urbana-Champaign
chang87@illinois.edu

Thomas Huang
Beckman Institute
University of Illinois at Urbana-Champaign
huang@ifp.uiuc.edu

Abstract

While geo-localization of web images has been widely studied, limited effort is devoted to that of web videos. Nevertheless, an accurate location inference approach specified on web videos is of fundamental importance, as it's occupying increasing proportions in web corpus. The key challenge comes from the lack of sufficient labels for model training. In this paper, we tackle this problem from a novel perspective, by "transferring" the large-scale web images with geographical tags to web videos, to make a carefully designed associations between visual content similarities. A group of experiments are conducted on a collected web image and video data set, where superior performance gains are reported over several alternatives.

Introduction

With the ever growing scale of social multimedia data, nowadays there are emerging opportunities along with significant challenges for understanding their contents to make a better sense of our visual world. One of the most significant trends is the so-called "geo-tagging", i.e., inferring the geographical locations from the visual content to "locate" information or data on the map. Given such a functionality, many potential applications are there for instance landmark retrieval (Kennedy et al. 2007), visual tour guide (Zheng et al. 2009), landmark recommendation (Gao et al. 2010), geographic-aware image search (Hays and Efros 2008), and mobile devices localization (Ji et al. 2009).

Despite success in geo-tagging of user-contributed images, how to infer the geo-tags from web videos retains an open problem, which is, however, of fundamental importance due to the increasing proportion of geo-tagged images in web data corpus. The difficulties are three-fold. On device, currently most of mobile phones and digital cameras do not record the geo-information when shotting videos; On annotation, there are limited geo-tagged benchmarks to train efficient geo-tagging model; finally on visual statistics, even a single video clip is composed with various scenes, which dramatically increases the difficulty in content analysis.

We conquer the above challenges from a novel perspective, by learning effective visual models alternatively from the image domain. Nowadays, there are increasing amount of geo-tagged images with rich groundtruth labels available. Such massive labeled data inspires us to transfer these geo-tags from web images to web videos.

To this end, the key challenge retains the heterogeneous distribution between images and videos, due to the difference in the visual quality and semantics. We tackle this problem following our previous attempt (Liu et al. 2011) by a novel search-based transfer learning algorithm using an AdaBoost (Freund and Schapire 1995) like learning structure. We train the landmark wised geo-tagging model from Flickr labeled images, the ensemble of which forms a "multi-class" classification problem for a given web video. In addition, we incorporate the temporal consistency to further improve the inference accuracy.

The proposed algorithm is testified on the dataset collected from Flickr and YouTube, with comparisons to different alternatives[1]. The results show that our method achieves significant improvement over various baselines. Besides, we also show its application on geo-based web video browsing.

Localizing Web Videos from Flickr Images

Figure 1 shows the framework of the proposed method. We start from learning accurate visual appearance models at the landmark level, which is achieved by modeling its individual viewing angles from Flickr images with geo-tag metadata. We then transfer this model to suit for the video based geo-tagging, by making an effective association between their heterogeneous visual similarity.

Building the Geo-Model from Flickr Images

Given a large scale image collection for a given landmark (as organized from its contextual tags or from clustering their geo-locations), we start from evaluating the trustworthy of Flickr images, as used subsequently to assess how

[1]To the best of our knowledge, this is the first work on this topic, and we only compare with some alternatives of our methods.

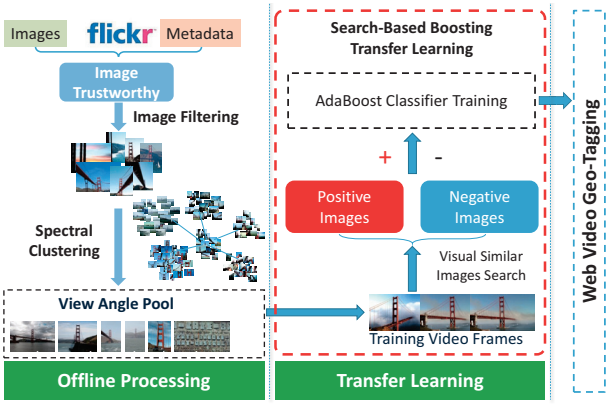

Figure 1: The framework overview of the proposed method.

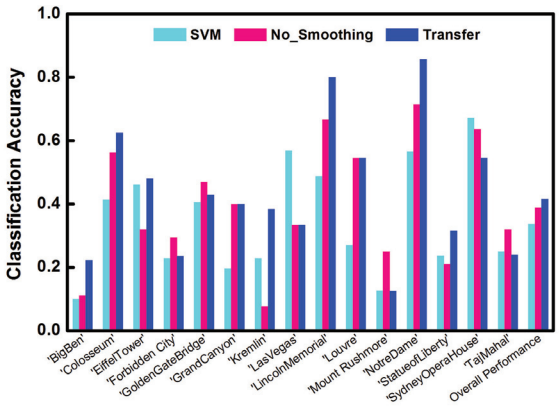

Figure 2: Classification performance comparisons for each landmark category

good an image is to build landmark appearance models for video geo-tagging. In this step, only "high-quality" images are selected for the subsequent process.

To this effect, considering even the images of the same landmark might be diverse in their view angles, we perform the spectral analysis (Ng, Jordan, and Weiss 2002) to divide the image collection I^l of landmark l into different view angles $\{I^{l,v_l}\}$, according to their visual similarity. As a result the image collection is organized as:

$$\mathcal{I} = \{I^{l,v_l} : l \in \mathbb{L}, v_l \in \mathbb{V}_l\}, \quad (1)$$

where $\mathbb{L}$ is the set of all landmarks, $\mathbb{V}_l, l \in \mathbb{L}$ is the set of all view angles from landmark l in our collection.

Furthermore, we use the visual representation in terms of SIFT points(Lowe 2004) of I^{l,v_l} to describe the model of each landmark at a certain view angle.

Search-Based Transfer Learning

Given the heterogeneous feature spaces between images and videos, the model learnt above cannot be directly applied to web videos. To reduce this cross-domain divergence as much as possible, we only transfer the "creditable" Flickr information to YouTube videos. For each training video frame f_i, the top k near-duplicated Flickr images, i.e., $\mathcal{D}_i = \{x_{i,j} : j = 1, ..., k\}$, are preserved as the training data. And the training of video geo-tagging model follows the discriminative principle: the classifier should not only separate the obvious positive (P: positives in $\mathcal{D}_i$ in our case) and obvious negative (N_d: the images of other landmarks and not in $\mathcal{D}_i$) samples, but also provide the discrimination on the ambiguous data (N_a: the negative images in $\mathcal{D}_i$). This principle is achieved by adopting an AdaBoost (Freund and Schapire 1995), as to penalize the ambiguous images N_a with $\eta > 1$ in the initialization.

More specifically, we penalize more on the ambiguous training data, and trust more on the near-duplicate Flickr images. It starts with an initialization of weighted training samples $\{(P, 1) \cup N_d, 1 \cup (N_a, \eta)\}$, following by training an AdaBoost classifier. As for the visual features classifiers training, we use SIFT (Lowe 2004) features and 2,000 dimensional *Bag-of-Words*.

In implementation, to fully explore the visual divergence between different view angles, we train the classifiers $\mathcal{G} = \{g^{l,v_l}\}$ for each view angle respectively, instead of on the landmark level. We then ensemble these classifiers into the classifier $g^l(x)$ of landmark l for the input video x:

$$g^l(x) = \max_{v_k \in V^l} g_k(x), \quad (2)$$

The temporal consistency of video frames are further incorporated into the above classifier, by performing a temporal Gaussian Filter to remove the incorrect geo-tags.

Experiments

We crawled a data set consists of 56,456 Flickr images and 2,000 web video clips from 15 landmarks. For each landmark, 8 view angles are clustered. We compared the proposed method with alternatives including the image classification using SVM directly trained on Flickr images (denoted as SVM), and classifier learnt using the proposed transfer learning without Gaussian smooth (denoted as No_Smoothing). The overall performance of SVM, No_Smoothing, and our method are 33.67%, 38.82% and 41.57% respectively, as shown in Figure 2.

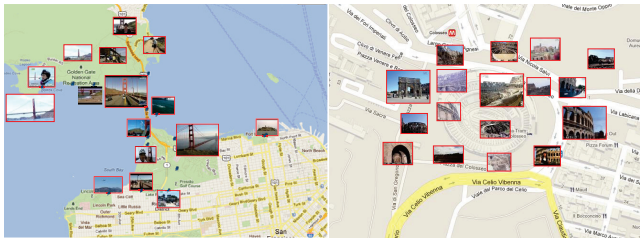

Figure 3: Application: Placing the web videos on the map.

A potential application is also demonstrated in Figure 3, in which we place the localized web videos [2] on maps using Google Map API, to show the potential of our method for location sensitive video search.

[2]We determine the location of each video according to view angle it belongs to, which is localized by averaging the geo-locations.

Acknowledgement

This work is supported in part by HP Innovation Research Program with UIUC, the 985 Project of Xiamen University, and the Fundamental Research Funds for the Central Universities of Xiamen University.

References

Freund, Y., and Schapire, R. 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, 23–37. Springer.

Gao, Y.; Tang, J.; Hong, R.; Dai, Q.; Chua, T.-S.; and Jain, R. 2010. W2go: a travel guidance system by automatic landmark ranking. In *Proceedings of the international conference on Multimedia*, 123–132. ACM.

Hays, J., and Efros, A. A. 2008. im2gps: estimating geographic information from a single image. In *CVPR*. IEEE.

Ji, R.; Xie, X.; Yao, H.; and Ma, W. 2009. Mining city landmarks from blogs by graph modeling. In *ACM Multimedia*, 105–114.

Kennedy, L.; Naaman, M.; Ahern, S.; Nair, R.; and Rattenbury, T. 2007. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *ACM Multimedia*, 631–640.

Liu, X.; Yao, H.; Ji, R.; Xu, P.; Sun, X.; and Tian, Q. 2011. Learning heterogeneous data for hierarchical web video classification. In *ACM Multimedia*, 433–442.

Lowe, D. 2004. Distinctive image features from scale-invariant keypoints. *IJCV* 60(2):91–110.

Ng, A.; Jordan, M.; and Weiss, Y. 2002. On spectral clustering: Analysis and an algorithm. *NIPS* 2:849–856.

Zheng, Y.; Zhao, M.; Song, Y.; Adam, H.; Buddemeier, U.; Bissacco, A.; Brucher, F.; Chua, T.; and Neven, H. 2009. Tour the world: building a web-scale landmark recognition engine. In *CVPR*, 1085–1092. IEEE.