Localizing web videos using social images

Liujuan Cao a, Xian-Ming Liu b, Wei Liu c, Rongrong Ji a,⇑, Thomas Huang b

a School of Information Science and Engineering, Xiamen University, PR China
b Beckman Institute, University of Illinois at Urbana-Champaign, USA
c IBM T.J. Watson Research Center, USA

Article history:
Received 23 February 2013
Received in revised form 23 July 2014
Accepted 3 August 2014
Available online 16 September 2014

Keywords:
Web video analysis
Cross-domain
Social media
Landmark recognition
Classification

Abstract

While inferring the geo-locations of web images has been widely studied, there is limited work engaging in geo-location inference of web videos due to inadequate labeled samples available for training. However, such a geographical localization functionality is of great importance to help existing video sharing websites provide location-aware services, such as location-based video browsing, video geo-tag recommendation, and location sensitive video search on mobile devices. In this paper, we address the problem of localizing web videos through transferring large-scale web images with geographic tags to web videos, where near-duplicate detection between images and video frames is conducted to link the visually relevant web images and videos. To perform our approach, we choose the trustworthy web images by evaluating the consistency between the visual features and associated metadata of the collected images, therefore eliminating the noisy images. In doing so, a novel transfer learning algorithm is proposed to align the landmark prototypes across both domains of images and video frames, leading to a reliable prediction of the geo-locations of web videos. A group of experiments are carried out on two datasets which collect Flickr images and YouTube videos crawled from the Web. The experimental results demonstrate the effectiveness of our video geo-location inference approach which outperforms several competing approaches using the traditional frame-level video geo-location inference.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

With the explosive growth of social networks, massive and heterogeneous multimedia data available in the Web provides a unique opportunity to bridge digital corpus and our physical world, which has become the mainstream of the ongoing multimedia research. To facilitate accurate recommendations by exploiting social media knowledge, web media tagging is also becoming an emerging important research direction [27,26,25,28,9].

Despite the widely studied semantic tagging scenario, annotating geographical locations of social media recently arises to be a major trend, which includes plenty of novel research subjects such as landmark retrieval and recognition, visual tour guide, geography-aware image search, mobile device localization, and aiding virtual reality.

Geographically tagging user-contributed images has been investigated in recent years [16,13]. However, tagging web videos has been less explored so far, in which inferring geo-locations of web videos is especially vital to location-based video services. The key difficulty of video geo-location inference is twofold. On the device level, current mobile phones and digital
cameras typically do not record the geo-information when shooting videos; on the other hand, there are limited geographic tags available for web videos, so it may be infeasible to train accurate classifiers or search engines in order to geographically tag the web videos.

The previous attempts of web video geo-location inference almost resorted to using metadata of social networks, such as titles, descriptions, and comments [17]. To improve the inference accuracy, extra evidence, e.g., visual and acoustic features, is incorporated [15]. However, limited work has successfully exploited visual content to assist the geo-location inference of web videos, mainly due to insufficient training examples that cause a challenge in effectively modeling the landmark appearances. This challenge originates from two aspects: (1) the low quality of web videos, which results in a limited amount of SIFT features being extracted and therefore compromises the accuracy of near-duplicate visual content matching; (2) the difficulty to diversify different scenes in the same video, which prevents assigning the correct geographic tags to the corresponding locations or physical scenes. We illustrate the challenge in Fig. 1.

In this paper, we propose to tackle this challenge from a novel transfer learning perspective, i.e., transferring an accurate and easy-to-learn video geo-tagging model from the image domain. Nowadays, there is an increasing amount of geo-tagged images available on the Web. Such massive image data prompts us to “transfer” the geo-tags from web images to web videos.

To do knowledge transfer across the image and video domains, we need to address two major issues: first, the tags of web images are usually noisy; second, the visual features of images and videos appear quite different due to their large variations. We address the first issue by proposing a web image trustworthy measurement to remove the “untrustworthy” web images. Afterwards, we perform a view-specific spectral clustering over the images of a given landmark to diversify different “views” of a single location. To build an effective video location inference model, a novel search-based transfer learning algorithm is proposed by constructing an AdaBoost [6] classifier in each view, and the outputs of multi-view AdaBoost classifiers are then combined into an overall landmark recognition model. In addition, we incorporate the temporal consistency to further improve the inference accuracy, which leverages the fact that temporally related video frames within the same video shot are more likely to be captured from the same view of a landmark or location.

To verify the proposed video geo-location inference model, we collect more than 50,000 geo-tagged images from Flickr and 2000 video clips from YouTube, respectively. Experimental comparisons on the two datasets show that our method achieves remarkable improvements over various competing methods. Besides, the proposed method can easily be integrated into the applications involving geo-location based web video browsing.

The rest of this paper is organized as follows: Section 2 briefly introduces the related work. The overall framework of the proposed method is presented in Section 3. Section 4 builds the effective landmark model using social images, which is further transferred to the video domain as described in Section 5. We conduct the experimental validations in Section 6. Finally, we conclude the paper and discuss the future work in Section 7.

### 2. Related work

Geography related image analysis has attracted intensive research attention from both academia and industry. One of the pioneering works comes from Kennedy et al. [16], which analyzed geographic tags (e.g., landmarks) of millions of Flickr images.

---

1. [www.flickr.com](http://www.flickr.com).
2. [www.youtube.com](http://www.youtube.com).
images and built an appearance model for each landmark by incorporating both metadata and visual features of images. In mobile devices, Microsoft Photo2Search [13] implemented a geo-location sensitive image search system to facilitate visual information driven localization and recommendation. Img2GPS [11] provided a novel perspective for image localization by estimating accurate geo-locations through statistical inference. Other related work include landmark search engines [33,20,31], location search [12,24,2,32], and the applications in mobile devices [1].

Our work also falls into the area of cross-domain transfer learning. Transfer learning makes use of auxiliary data information from a source domain to make predictions in a heterogeneous target domain [22], which is widely adopted in web multimedia analysis. For instance, in the task of video concept detection, classifiers are initially trained in a source domain (e.g., news videos and sports videos), and further applied to a target domain (e.g., consumer videos and surveillance videos) [30,14,3]. With the explosive increase of web media, collaborative information can be leveraged into web video analysis. Duan et al. made use of YouTube videos to train the cross-domain classifiers for consumer videos [4], while Jiang et al. incorporated heterogeneous Flickr images to estimate the taxonomy which is used in video concept detection [14]. Most recently, Liu et al. connected heterogeneous web media sources by associating Flickr images and YouTube videos according to the visual and textual similarities [18].

The key contribution of our work is to exploit the auxiliary cross-domain information to learn an effective geo-tagging model for web videos. To the best of our knowledge, this is the first attempt to employ cross-domain information for web video localization. Our proposed approach overcomes the limitation of insufficient credible training examples, bridging the visual-geographic information across heterogeneous web media sources.

3. Framework

The framework overview of the proposed method is displayed in Fig. 2. In this paper, our main efforts lie in the following two aspects: mining a visual model from Flickr images (the offline image processing component), and effective transfer learning from the Flickr image domain to the YouTube video domain (the transfer learning component).

To learn a reliable geo-location inference model using the Flickr images and their metadata, we first propose an image trustworthy measurement regarding the consistency in visual features and metadata of the web images motivated by [16]. Such a measurement is able to eliminate the noisy images whose visual contents are irrelevant to the associated geographic tags. To better utilize the purified geo-tagged images, we adopt a two-step clustering procedure: in the first step, all the images are clustered at the landmark level; in the second step, a spectral clustering is executed to separate the landmarks to different views based on their visual disparities. This leads us to build the view-level landmark models.

We further adapt the learned geo-location inference model to the video domain by applying search-based transfer learning. We model the image-to-video domain transfer problem as a re-sampling process in training, which is implemented by using the AdaBoost [6] algorithm. A classifier for each view is trained beforehand, and then the classification results are assembled to form the transferred geo-tagging model. Finally, in the online geo-tagging part, each input web video is classified by temporally integrating the frame-level geo-tags using the obtained view-specific classifiers, which avoids the misclassification of single video frame.

Fig. 2. The framework overview of the proposed method which is composed of three main components: offline image processing, image-video transfer learning, and web video geo-tagging.
4. Building the geo-model from Flickr images

Given a set of Flickr images associated with a specific geographic tag like Statue of Liberty or Time Square, in this section, we first address the problem of noisy image filtering by proposing an image trustworthy measurement. Then, the corresponding geo-model is built with a data-driven approach. We evaluate the image trustworthiness for every image in the dataset according to the attached metadata in Section 4.2. While in Section 4.3, the images are firstly clustered into landmark-level image sets according to their GPS locations, followed by a visual appearance clustering step to further group those images within a given landmark into different views.

4.1. Preliminary: visual feature extraction and similarity measurement

To represent the visual content of images, we use both global and local visual features. For the global features, Color Moment and Gabor Texture are adopted. While for the local features, the SIFT [19] descriptor is used. Moreover, we follow the approach in [23] to quantize the SIFT features extracted from each image into a Bag-of-Words histogram by using a dictionary of 2000 visual words.

For each pair of images $l_i$ and $l_j$, the similarity is defined as the linear combination of both global features and local features. In specific, the similarity between two images is defined as

$$s_{ij} = \lambda_1 \frac{1}{d_{\text{color}}(i,j)} + \lambda_2 \frac{1}{d_{\text{gabor}}(i,j)} + \lambda_3 \frac{1}{d_{\text{BoW}}(i,j)},$$

where we use the Euclidean distance to measure the color and texture similarities, and use the Manhattan distance to measure the Bag-of-Words similarity. All the distances are normalized to the range of [0, 1], and we use equal weights (i.e., $\lambda_1, \lambda_2, \lambda_3$) for all the distances.

4.2. Image trustworthy measurement

The trustworthiness of web images is leveraged to assess how “suitable” an image is for the subsequent analysis, which aims for removing the noisy (or “bad”) images from the image collection. The evaluation measurement of the trustworthiness makes use of both the geo-information and other metadata (such as user comments, view times, and favorites) to identify to what extent we can trust and use a specific image in our inference.

More specifically, we formulate the image trustworthiness $T_i$ for image $i$ as

$$T_i = \alpha F_i + (1 - \alpha) D_i,$$

where $\alpha \in (0, 1)$ is a constant, $F_i$ is the percentage of positive response (including comments, like, views, etc.), and $D_i$ measures the divergence of geo-location for image $i$ in terms of landmark $l$, which is

$$D_i = \frac{1}{(x_i - \mu_l)' \Sigma_l^{-1} (x_i - \mu_l)},$$

where $x_i$ is the visual feature of image $i$ while $\mu_l$ is the mean vector of all the images labeled as landmark $l$. Hence, $D_i$ is the weighted divergence by using the covariance of each landmark $\Sigma_l$.

The principle behind Eq. (2) is that we usually trust the images:

1. with more positive user feedbacks, such as more people liking these images, and more user comments;
2. the visual features distribute commonly with the other images of the same landmark in the feature space, which is measured by Mahalanobis distance (Eq. (3)) between the image and the averaged center of a specific landmark; the usage of Mahalanobis distance is to reduce the bias of image features at the landmark level.

In our scenario, we keep the web images with the largest trustworthiness, and consequently remove the other images that are regarded noisy.

4.3. The geo-model

Given a Flickr image collection $I$, our purpose is to represent each landmark with a collection of related images and their visual features in a data driven manner. However, considering the situation that even the images of the same landmark with almost identical GPS locations may be diverse in the view angles, we then prefer to further partition the image collection in the view level as follows:

$$I = \{l^n_i : l \in L, v_i \in V_l\},$$

where $l$ denotes the landmark, and $v_i$ denotes the view angle for $l$. To achieve this goal, we first cluster the images into $K$ landmark clusters $I = \{l^n_i\}_{K=1}^K$ according to their tags and GPS locations. Then, we follow the classical spectral clustering
algorithm \[21\] to divide the sub-image collection \( I_l \) into different view angles \( I_l = \{I_l^i\} \) according to the image similarity of visual features, which is already defined by Eq. (1). Consequently, the image similarity matrix is constructed as \( S = [s_{ij}] \), followed by a Laplacian normalization:

\[
W = I - D^{-1/2}SD^{-1/2}, \quad D = \text{diag}(d_{ii}), \quad d_{ii} = \sum_j s_{ij}.
\]

Working on the eigen-spectrum of the matrix \( W \), the spectral clustering algorithm \[21\] is performed to group each sub-image collection in the landmark level into the view-level sub-clusters. An example of the spectral clustering result for the landmark “Golden State Bridge” is shown in Fig. 3, which indicates that the proposed pre-processing can well separate different view angles into different sub-clusters.

5. Search-based transfer learning

Due to the gap between the cross-domain feature spaces, the model learned from Flickr images cannot be directly applied to the YouTube videos. Typically, the Flickr images and YouTube videos are with different resolutions and photographing quality. To solve this problem, we propose a search-based transfer learning algorithm in this section.

5.1. Transfer learning

To reduce the cross-domain variance as much as possible, we only transfer the “creditable” information from Flickr images to YouTube videos. Motivated by the work of “Annotation by Searching” \[29\], we define the “creditable” as the near-duplicate Flickr images with the training video frames, which is further used as the training data for the web video classifiers.

For each training video frame \( f_i \), the top \( k \) near-duplicate Flickr images based on global features, \( D_i = \{x_{ij} : j = 1, \ldots, k\} \), are found as the training data. As for the global features, first each image and video frames are downsamples to 128 x 128 pixels, to get rid of the negative effect caused by different resolution. Then we extract the Color Histogram as the main feature used in the near-duplicate search.

The training of the geo-tagging models for videos follows a discriminative principle: the classifier should not only separate the true positive (i.e., the near-duplicate Flickr images) and negative (i.e., out of the near-duplicate Flickr images) samples, but also have the ability to distinguish the ambiguous data (i.e., the negative images within the returned duplicate
images). To achieve this goal, we introduce an transferring algorithm that borrows the idea from AdaBoost as shown in Algorithm 1.

**Algorithm 1: Search-Based Transfer Learning Algorithm**

**Input:** Training video frame collections $V^l = \{f^l\}$, Flickr Images $\mathcal{I} = \{I^{lm}\}$

**Output:** classifier $g(\cdot)$

1. $P = \emptyset, N_d = \phi, N_a = \phi$

2. foreach video frame $f_i^l$ do

3. Search $k$-size near-duplicate images $\mathcal{D}_i = \{x_{i,j} : j = 1, ..., k\} \subseteq \mathcal{I}$;

4. Add the negative samples in $\mathcal{D}_i$ into $N_a$, as:

5. $N_a = N_a \cup \{x_{i,j} : x_{i,j} \in \mathcal{D}_i \cap \{f\}\}$;

6. Add the positive samples in $\mathcal{D}_i$ into $P$, as:

7. $P = P \cup \{x_{i,j} : x_{i,j} \in \mathcal{D}_i \cap \{f\}\}$;

8. end

9. Construct $N_d$: random sample $N_d = \{x_i | x_i \notin f\}$;

10. foreach $x \in N_d$ do

11. weight the training sample: $w_x = 1.0$;

12. end

13. foreach $x \in N_a$ do

14. weight the training sample: $w_x = \eta$ where $\eta > 1$;

15. end

16. Train the classifier $g(x_w)$ using AdaBoost[6] on:

17. $P, (N_d, W_d), (N_a, W_a)$

18. return $g(\cdot)$

The key of Algorithm 1 is to penalize more (weighted $\eta > 1.0$) on the ambiguous training data, and trust more on the global near-duplicate Flickr images. More specifically, by using the proposed feature space transfer, the algorithm should not only correctly classify the aligned training samples, but also emphasize more on distinguishing the ambiguous data. As for the visual features used for classifier training, we use the same features introduced in Section 4, that is, SIFT features and 2000-dimensional BoW.

In the implementation, to fully explore the visual divergence between different viewing angles, we train the viewing angle level classifiers $G = \{g^{lm}\}$ to replace the original landmark level classifier.

5.2. Localizing the web videos

To localize the web videos, we use the pre-trained view classifiers $G$ following an ensembled manner: for a given Landmark $l$, its classifier $g^l$ is obtained by utilizing:

$$g^l(x) = \max_{n \in V^l} g_n(x),$$

(7)

where $V^l$ is the set of view angles for $l$, and $x$ is the visual feature vector of the video frame to be classified. Different from the training procedure in which the near-duplicate Flickr reference images are involved, the classification process deal with the video frame directly. Thus, the time cost of classification is linear to the number of weak classifiers used in the algorithm.

Another important property of video sequence is the temporal consistency, which restricts that frames within the same shot will be temporal consistent. Thus, we uniformly sample video frames from input videos, and classify each frame using the classifier in Eq. (7), and perform a temporal Gaussian Low-Pass Filter to the geo-tagged video sequence. In practice, we adapt the Gaussian window size according to the shot length (with factor = 0.6), and a fixed kernel width $\sigma = 1.0$.

For each sampled frame, the geographical model with the highest classification score will be chosen as its geo-tag, and for each video shot, its geo-tag is determined by the maximum modal number of all its frames.
6. Experimental results

In this section, we will first introduce the used datasets and experimental settings. Then we will present in details a group of experiments conducted on the collected datasets. Finally, we will further demonstrate an application of our approach, which targets at placing web videos on a geographic map.

6.1. Preliminary

In order to gather sufficient training data, we downloaded the web images from Flickr with tags being 15 cities as follows: “Big Ben”, “Colosseum”, “Eiffel Tower”, “Forbidden City”, “Golden Gate Bridge”, “Grand Canyon”, “Kremlin”, “Las Vegas”, “Lincoln Memorial”, “Louvre”, “Mount Rushmore”, “Notre Dame”, “Statue of Liberty”, “Sydney Opera House”, and “Taj Mahal”. Totally, we have collected 56,456 web images that are used for training. For each landmark, 8 top view angles with the most representative images are preserved. In summary, we have collected 120 view angles from 15 landmarks, which make training 120-fold weak classifiers for location recognition.

The videos downloaded from YouTube are partitioned into a training set and a test set. Among the downloaded 2000 video shots, we adopt 50% for running our transfer learning algorithm, and the remaining 50% for testing. The video frames are extracted from these videos in a gap of 3 s. Fig. 4 shows some sample video frames used in this paper.

6.2. Parameter validation

An important parameter needed by our algorithm is the number $K$ of the near-duplicate Flickr images appearing in Algorithm 1. We test our algorithm on the 50,000 Flickr images by varying $K$ from 100 to 1000, with the results shown in Fig. 5.

The results in Fig. 5 indicate that our algorithm can obtain more input training data which is very helpful to improving the robustness of the learned classifiers. However, on the other hand, due to the limited scale of the collected Flickr image dataset, a large $K$ tends to introduce ambiguous negative samples in the training stage, which therefore may lead to biased classification results. More specifically, as $K$ increases, the performance first increases and then decreases. It is worth pointing out that the best performance is achieved around $K = 300$ on the Flickr dataset.

Moreover, the optimal value of $K$ should depend on the size of the training samples, and the changes from landmarks to view angles. As a result, there exists no general answer to what is the best $K$. However, since we pay more attention to the overall performance, for simplicity we choose $K = 300$ in our experiments, which is optimal on the current Flickr dataset.

6.3. Results and discussion

In this section, we verify the proposed approach in terms of the task of inferring the geo-locations of web videos. Since there is no comparable method for web video localization using social images such as we did in the paper, we only compare our method (denoted as “Transfer”) with different scenarios mentioned in the paper.

Fig. 4. Sample video frames used in this paper. The videos vary in view angles, environment changes, and lighting conditions.
1. Directly training the SVM classifiers using LibLinear [5] over the downloaded Flickr images (denoted as SVM); the Gaussian smoothing is not adopted in this setting.
2. Learning the classifiers using the proposed transfer learning algorithm but without Gaussian smoothing over the video frames (denoted as No Smoothing).

The overall performance of SVM, “No Smoothing”, and our method (”Transfer”) are 33.67%, 38.82%, and 41.57%, respectively. The detailed comparisons for different landmarks are shown in Fig. 7. It is obvious that in most cases the performance is improved dramatically, especially for the landmarks “Lincoln Memorial”, “Louvre”, and “Grand Canyon”. To be noticed, the performance for the landmark “Las Vegas” drops greatly. This is because the visual appearances of “Las Vegas” vary largely, which poses a big challenges for transfer learning to learn effective classifiers.

Some examples of good and bad classification results are also shown in Fig. 6. Fig. 6(a) shows the localization results of our proposed method, from which the tolerance against view angle changes is revealed. And according to Fig. 6(b), most false classifications are due to the fact that their ambiguous representations of visual features, i.e., using only visual features is not discriminative enough to achieve accurate localizations of web videos. An obvious example is the case that the landmark “Statue of Liberty” is incorrectly inferred as “Lincoln Memorial”, since few SIFT feature points can be extracted in the video frames. We also realize the need to utilize the metadata of web videos in the future.

To sum up, the improvements in the experimental results have revealed that:

• It is helpful to leverage extra sources of information to assist geo-tagging. In the experiments, we only use a limited number of training videos, while our improvements over the competing methods are already promising. This provides a feasible solution to various similar problems.
• The proposed transfer learning algorithm is effective to handle the domain adaption.
• Distinguishing view angles is essential to dealing with view-point changing in real systems, according to the results shown in Fig. 6.

6.3.1. Potential application

Two examples of different locations for an application are also demonstrated in this section, as shown in Fig. 8. We place the localized web videos on the maps using Google Map API, to show the potential of our method being applied to location-sensitive video search.
7. Conclusion and future work

This paper serves as the first work endeavoring to infer the geographical locations of web videos. The core technique of our approach is directly transferring the geo-tagging knowledge from web images to web videos, without training any video-domain geographical inference model that suffers from inadequate labeled data. There are several key components in our proposed novel geo-location inference framework, including a spectral clustering algorithm to diversify multiple views of web images for a given landmark, an ensemble learning algorithm to deal with the degenerated video content, and a temporal window smoothing scheme to yield temporal-consistent video frames. We have testified the proposed approach on two collected datasets consisting of more than 50,000 web images and 2000 web video clips crawled from the Web. Our approach showed superior performance over the alternatives.

As for the future work, one direction is to improve the inference performance by using better visual feature representation. Some approaches such as combining multiple modalities of features [10] could remove the ambiguity in visual information. In order to develop more robust learning models for the cross-domain problem studied in the paper, a feasible way is to impose more reasonable assumptions upon the data distributions, or explore the underlying structures of the data distributions like [26]. Regarding the recent success in image retrieval with 3-D object modeling [8,7], it also makes sense to estimate the 3-D landmark parameters for building appropriate 3-D models.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant 61422210 and Grant 61373076, the Fundamental Research Funds for the Central Universities under Grant 2013121026, and Xiamen University 985 Project.