Towards Lifestyle Understanding: Predicting Home and Vacation Locations from User's Online Photo Collections

Danning Zheng, Tianran Hu, Quanzeng You, Henry Kautz and **Jiebo Luo**
University of Rochester
Rochester, NY 14627
dzheng2@u.rochester.edu, {thu, qyou, kautz, jluo}@cs.rochester.com

Abstract

Semantic place labeling has been actively studied in the past few years due to its importance in understanding human mobility and lifestyle patterns. In the last decade, the rapid growth of geotagged multimedia data from online social networks provides a valuable opportunity to predict people's POI locations from temporal, spatial and visual cues. Among the massive amount of social media data, one important type of data is the geotagged web images from image-sharing websites. In this paper, we develop a reliable photo classifier based on the Convolutional Neutral Networks to classify the photo-taking scene of real-life photos. We then present a novel approach to home location and vacation locations prediction by fusing together the visual content of photos and the spatiotemporal features of people's mobility patterns. Using a well-trained classifier, we showed that the robust fusion of visual and spatiotemporal features achieves significant accuracy improvement over each of the features alone for both home and vacation detection.

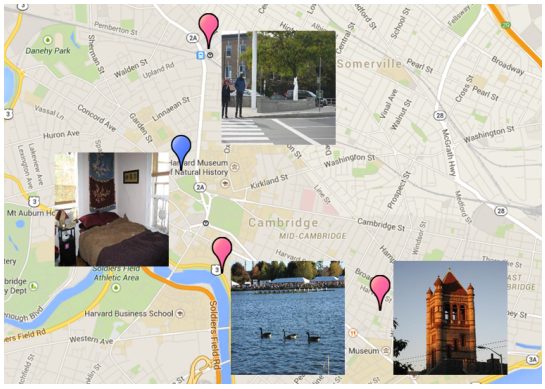

Figure 1: Visualization of a Flickr user's activity trace in Boston. The four pins represent the top 4 most frequently-visited locations, with the home colored as blue and non-home locations colored as pink. Each pin is shown with a photo taken at that location.

Introduction

Personalized semantic POI labeling is drawing much attention recently because of the huge impact it could bring to the study of human lifestyle, urban planning, and so on. In such a problem, we need to predict a label for the POIs of one's trace. Different from the typical location labeling or classification problem, personalized semantic POI labeling considers the various meanings of a single place to different people. For example, person A takes a vacation at the beach and person B works at the same beach. In this scenario, to A the semantic label of the beach should be "vacation" while to B the label should be "work". In this paper we propose a machine learning method to semantically label two important POIs in one's daily life – home and vacation.

Precise home location is increasingly important in various researching fields. In urban planning, knowing location-based behavior can help build more optimal design of urban environment, including the transportation networks and pollution management. Research areas such as disease propagation and outbreak modeling all require the knowledge on where people live. In addition, home plays the role of the origin of daily life for most people; it provides a reference point to other semantically meaningful places. In other words, home information helps the prediction of other POIs. For instance, given the home location of a person then the places which are quite far away from it are unlikely to be his/her work place.

Because of the importance of home location to people's mobility patterns, in this paper we first work on predicting the location of people's homes. Based on our precise and accurate home location inference we further predict another important POI in people's activity trace – vacation locations.

Existing methods that can precisely detect home location are all based on surveys, GPS data or cellular telephone records (Krumm and Rouhana 2013; Hoh et al. 2006; Cho, Myers, and Leskovec 2011). However, the process of obtaining such continuous data are often resource demanding and not scalable. Also, due to the limitation of the dataset, GPS data and surveys are often not adaptable for follow-up studies. For example, although the American Time Use Survey (ATUS) provides comprehensive records of ATUS respondents' activity traces and demographic information, such information is not adaptable for follow-up investigations since we cannot correlate the user informa-

tion with any other data sources. In contrast, the availability of vast amounts of geotagged data available on social networks enables a low-cost and more flexible way to detect home location. Previously, researchers have built models to infer the home location of a person based on his or her online activities such as tweeting (Cheng, Caverlee, and Lee 2010) or check-ins (Pontes et al. 2012b). One of the main existing issues is that these methods either suffer from coarse granularity, at only city (Pontes et al. 2012a) or state level, or result in a low accuracy, at around 50% (Cheng, Caverlee, and Lee 2010).

A picture is worth a thousand words. In this paper, we address the home prediction problem by analyzing photos mined from Flickr. As a popular online photo-hosting community, Flickr has more than 3.5 million new images uploaded per day (Jeffries 2013). We apply machine learning techniques to geotagged Flickr images and automatically predict a Flickr user's home location within a 100-meter by 100-meter square on the basis of his or her posted images. Based on the home location, we further extract the features of locations for vacation prediction, including the distance from a location to one's predicted home location. Using these features, we train another model to automatically label vacation places for each user.

Our results has shown that the visual content of images can provide valuable clues complementary to the metadata captured with photos and can be used to improve personalized semantic POI labelling.

The contributions made in this study are threefold. First, we develop a reliable classifier by the Convolutional Neutral Network (Krizhevsky, Sutskever, and Hinton 2012), which can recognize the photo-taking scene of real-life photos. Second, we fuse the visual content of user photos with the spatiotemporal features of a user's activity to construct a robust multi-source home predictor, where each of the two modalities contributes to the improvement in home location. The precision to which we can locate a person allows various location-related research in finer granularity and with higher accuracy. Third, based on the predicted home location, we further propose a machine learning framework to identify the vacation locations.

Related Work

Locations such as home, working places and restaurants are important in understanding human mobility patterns and automatically predicting future activities.

Using the GPS data collected from users' vehicle, Krumm et. al (2013) extracted several temporal and spatial features and developed a rule-based classifier to predict one's home location. In their approach, it turns out the feature 'last location of a day' is the most significant feature in home detection. Also using GPS data, Liao et. al (2005a; 2005b) proposed a machine learning approach based on MCMC to identify a user's significant POIs as well as different activities taking place at the same location.

Besides taking advantage of GPS data to semantically label the locations of one's trace, Krumm et al. (2013) developed a machine learning algorithm to classify locations into different categories based on ATUS, a diary survey containing detailed record on the amount of time and the location Americans spent doing various activities. They used demographic and temporal features of people's activities to infer a place's label and their results showed that home location can be predicted with a high accuracy of 92%.

As people spend more time online, social networks enable an alternative approach to semantically label geographic locations. Cheng et al. (2010) used a Twitter user's tweet content to predict his or her home city based on the idea that the frequency and dispersion of a specific word in tweets should be different across cities due to regional differences. By purely analyzing the content of a user's tweet, Cheng managed to place a user within 100 miles of his or her actual location with a 51% accuracy.

Our work is also closely related to the study of semantic annotation of web images (Luo et al. 2008; Yuan, Luo, and Wu 2010; Hays and Efros 2008; Cao et al. 2009; Zheng et al. 2014). As photographic devices with GPS capability become more prevalent in the market, the massive amount of web images serve as an alternative data type to predict home location. In the last few years, many computational approaches have been used to recognize objects of certain types (faces, water, cars, buildings) and the scene (park, residential area) in a photo. James et al. (2008) estimated the geographic location of an image based solely on its image content. In (Joshi and Luo 2008), Joshi et al. described a framework to model geographical information. Based on a series of geotagged photos, Yuan et al. (2010) detected the associated event by fusing visual content and the associated spatiotemporal traces. Their result substantiated that the visual content and GPS traces are complementary to each other, and a proper fusion can improve the overall performance of event recognition. Similarly, a photo taken by a personal camera and a satellite image are combined to help improve picture-taking environment recognition in (Luo et al. 2008).

Data

In this section, we describe how we obtained the dataset used to train and evaluate the home and vacation predictors.

Since most Flickr user profiles do not have detailed home address information, we could not build the ground truth on user profiles. Instead, we used the geotags of a user's taken photos to precisely locate his or her actual home. We selected a set of tags, including "home", "kitchen", "living room", "family dinner" and their variants, and refer to them as *home-related tags*. Note that we have manually checked the photos with these home-related tags to make sure that most of the returned photos are highly related to home. Using the Flickr API, we collected all geotagged photos with home-related tags in the following populated areas: *Chicago, Boston, Austin, Columbus, Washington DC, Denver, Houston, Los Angeles, Salt Lake City, the greater NYC Area, the Bay Area, Phoenix, San Antonio, and Seattle*. Each photo is associated with a geographic tag accurate to the street-level, which is represented by a pair of longitude and latitude coordinates. Next, we manually picked out the photos that are taken at home and used the associated geotags as

the actual home locations. Altogether, we have collected the home locations of 1000 users.

For each user i, we extracted from the photo metadata a sorted time sequence $t_i = \{t_{i1}, t_{i2}, ...\}$, where t_{ij} represents the time point of user i's jth photo taken over a significant period. In consideration of home relocation, we queried Flickr for all public photos posted by these 1000 users in a one-year period, which is obtained by adding and subtracting half-a-year from the median time point in sequence t_i. Together these users have taken 423047 geotagged photos in a year.

Home Prediction

Since our first goal is to predict users' home location, we only keep the photos that are taken within the bounding boxes of the fourteen areas we mentioned earlier as our training dataset. After this procedure, we are left with 183291 geotagged photos taken by 1000 users. We divided the map into 100-meter by 100-meter squares and represent each geographic location as the central point of the square it falls into. Therefore, if we can correctly predict the square, the distance error will be no more than 70.7 meters.

For each user, a photo taken with a geotag is considered as one check-in at that geographic location. In the Flickr dataset, there exists a certain amount of locations which are visited by more than one user, but a location can only be the home of one user. Therefore, in order to differentiate a location by users, we use a pair (i, j) as a sample ID to represent a location j being checked-in by user i. Altogether, we have recorded 31053 unique (*user*, *location*) sample IDs by 1000 users.

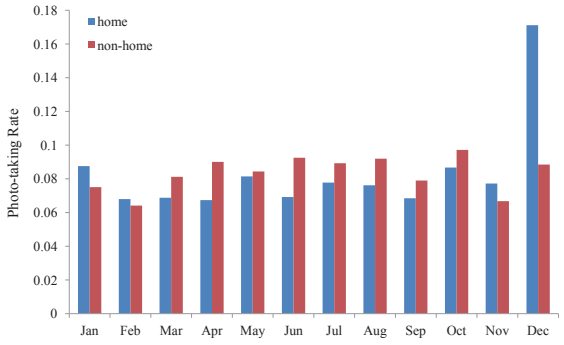

Figure 2: Comparison of the number of uploads at home/non-home locations on a monthly basis. Y-axis represents the percentage of the number of "home photo"/"non-home photo" uploaded during a specific month. Note the strength of home photos in December.

Temporal Features

According to previous work (Pontes et al. 2012b), home is supposed to be one of the most frequently visited places in a user's mobile trace. Therefore, we started by using the most frequently visited location as a preliminary prediction of a Flickr user's home location. We consider this

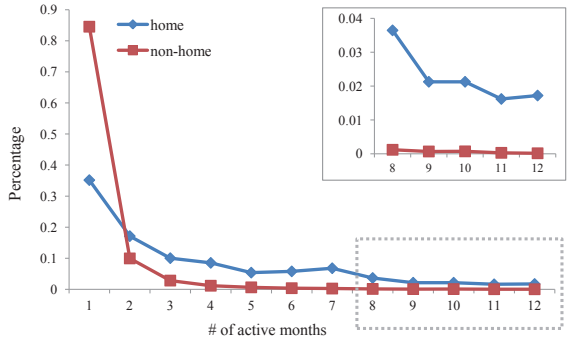

Figure 3: Comparison of the number of active months at home/non-home locations during a year. Y-axis represents the percentage of home/non-home locations that are active for a specific number of months. The plot on the top right corner is a magnification of the part in the dotted box. Our data shows that people rarely stay at non-home locations for more than one month.

model as the baseline and refer to it as the *most check-in* method. In addition to this baseline, we then mine a large collection of temporal features for each unique (*user*, *location*) sample. As validated in previous work (Ye et al. 2011; Gao et al. 2013), human mobile behavior displays strong temporal cyclic patterns and this temporal regularity can help improve the performance of location prediction. Finally, we explore the feasibility to automatically assign a semantic label to a photo. We test the effectiveness of photo clue by adding visual content feature to our collection of temporal features and compare the performance of home prediction.

Similar to previous work (Ye et al. 2011; Gao et al. 2013), the Flickr data set shows strong evidence of yearly patterns (months across a year) of a Flickr user's photo-taking activity. Figure 2 demonstrates a significant difference between the number of photos taken at home and non-home locations on a monthly basis. December stands out from all the months in the sense that the number of photos taken at home in December is significantly higher than that in other months. Numerically, among all photos taken at home, the number of photos taken in December accounts for nearly 20% of the total photos. Note that this phenomena is specific for home since the number of photos taken at non-home locations is almost evenly distributed over the 12 months. This distribution is probably because people spend more time at home with family during Christmas and take plenty of photos during that time.

Another important observation is that the photo-taking activity at home is more prevalent across time since people can take photo at home at any time during a day, any day during a week and any month during a year. We define a *(user i, location j)* pair as active during a time period $[t_1, t_2]$ if user i takes at least one photo during $[t_1, t_2]$ at location j. Based on this definition, we use the number of active months, active hours and active days (out of a week) to quantify this temporal prevalence feature. Taking month as an example, Fig-

ure 3 shows that the number of active months at home are universally larger than those at non-home locations. More than 50% of the home locations are active for at least three months while nearly 90% of the non-home locations are active for only 1 month. This distribution reveals that although people may take a massive amount of photos during certain events such as commencement, wedding and vacation, such events would only happen once or twice during a year.

Clearly, there exists a high correlation between time and the number of photos in the Flickr dataset. Therefore, we extract a large collection of temporal features to represent each unique (*user*, *location*) sample. Since the distribution of user uploads are highly skewed (75% of the photos are uploaded by 20% of the users), we use the upload rate instead of the absolute number of uploads. For example, the January upload rate for a *(user i, location j)* pair is given by:

$$\frac{\text{\# of uploads in January by user } i \text{ at location } j}{\text{total \# of uploads by user } i} \quad (1)$$

Altogether, for each (*user*, *location*) sample, we extracted 16 temporal features, including the check-in rate, monthly upload rate, # of active hours, # of active months and # of active days.

Visual Feature

Different from photo tags and descriptions, which are usually not available or informative enough, visual content is always available for each photo. As an inherent feature, visual content can provide us fundamental insight on where a photo was taken. For example, a photo of family party is highly probable to be taken at home. Therefore, to take advantage of the rich information embedded in photos, we trained a classifier to distinguish "home-like" photos from the others.

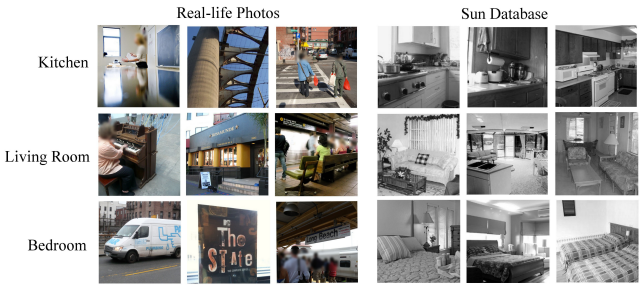

Figure 4: Examples of real-life and sun database photos classified as "kitchen", "living room" and "bedroom" by using HOG 2×2 features.

Scenes recognition approaches can be employed to extract the semantic content of pictures. In (Xiao et al. 2010), HOG2×2 features were used to classify photos into 397 categories (e.g. living room, kitchen) and achieved a higher accuracy than other single feature based methods. To distinguish photos taken at home from the others, we extracted a 300-dimensional HOG feature vector from each photo collected from Flickr. A well-trained SVM model was employed to classify the photos as home or non-home. Although the HOG feature works well on "clean photos" in which elements are obvious and well-constructed, the variability of real-life photos make it extremely challenging for classification. Real-life photos taken at home have various kinds of noise, with people and pets appearing in the photo as the most common one. In Figure 4, we show the classification result of HOG2×2+SVM. The classifier produces desirable results on the Sun database, but performed poorly when we applied it on real-life photos.

Inspired by some recent successes (Ross et al. 2014), we chose instead to employ a deep network to reliably assign semantic labels to a photo. For our purpose, each photo is classified as either taken at home ("home photo") or not taken at home ("non-home photo"). For each *(user, location)* sample, we define the "home photo" rate as:

$$\frac{\text{\# of home photos uploaded at location } j \text{ by user } i}{\text{total \# of home photos uploaded by user } i} \quad (2)$$

and use it as the visual content feature. As described in (Krizhevsky, Sutskever, and Hinton 2012), we extract a 4096-dimensional feature vector for each photo by using the Caffe (Jia 2013) implementation of the Convolutional Neutral Networks. Since our goal is to classify real-life photos, we chose to also use real-life photos as the training set to obtain optimal effect. We fine-tuned the pre-trained ImageNet model with an independent photo dataset consisting of 6000 home photos and 24000 non-home photos. All training photos are obtained by first querying Flickr for images with home-related tags and then manually checking to only keep photos that are taken in real life. Specifically, we filtered out the photos that look too standard, such as photos of model houses and hotels. We also purposely kept some photos taken at home with people or pets in the scene.

With the ground truth and the features mentioned above, we trained an Bayesian Network meta-classifier using the Weka toolkit (Witten and Frank 2005) over the set of (*user*, *location*) samples. Three different combinations of features: 1)temporal feature alone, 2)visual content feature alone, and 3)temporal+visual content feature, are examined and compared to the baseline method (most check-in). In our experiment, two-fold cross-validation is used to validate the robustness of our methods.

Experiments

In this section, we first present the result of home photo classification by CNN. The deep network is tested on all 47793 images scrawled from Flickr. Since it is impossible to label the whole dataset, we manually check the photo classification results to verify that the overall performance is reliable. We then evaluate the effectiveness of the proposed fusion of temporal and visual content features in predicting home location on the Flickr data. Prediction accuracy is used as the performance measure and is defined as:

$$\frac{\text{\# of correctly predicted users}}{\text{\# of total users}} \quad (3)$$

The second metric we use is the distance error. It represents the granularity level of home prediction and is defined as the distance from the geographic coordinate of the predicted home to that of the actual home. We compare the prediction

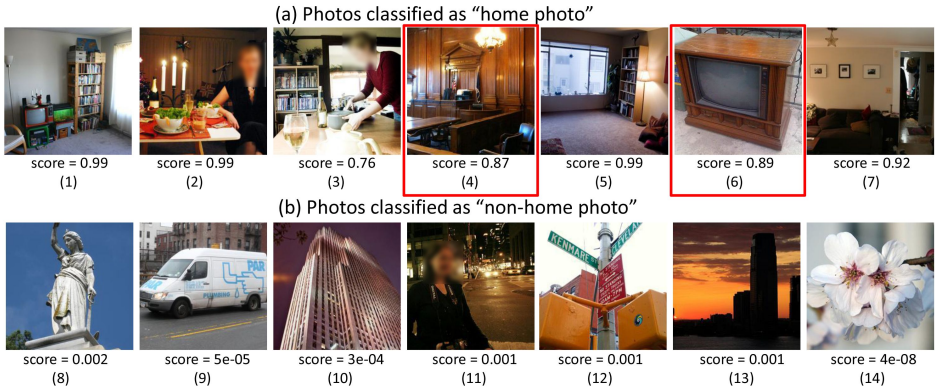

Figure 5: Examples of photos classified by the trained deep networks as (a) “home photo” and (b) “non-home photo”. Photos marked in red boxes are misclassified.

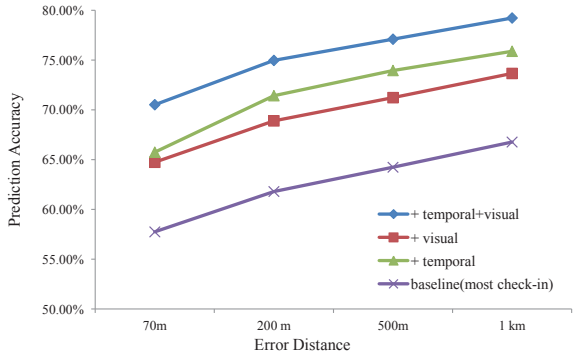

Figure 6: The performance of the baseline and the fusion home-predictor. The plot shows the prediction accuracy with increasing distance error tolerance (70 meters to 1000 meters).

accuracy of all four methods mentioned above with different distance error tolerance.

A few representative examples of photos are presented in Figure 5 to illustrate the performance of photo classification by CNN. Each photo is associated with an estimated score, which can be considered as the probability of being a “home photo”. The “home photo” examples show that the photo classifier can accurately identify certain home-related objects such as tables, shelves and sofas (photo #2, #5 and #7). However, some confusing scenes might be falsely classified as at home due to its similar structure or layout to a home. For example, the court (photo #4) and a discarded TV on the street (photo #6) are misclassified as at home. Overall, the main confusion comes from home-related objects or home-like structures, which are difficult to differentiate by a computational or even manual approach. The “non-home photo” examples reveal that the photo classifier can accurately identify outdoor photos even for portrait-oriented photos. Comparing photo #3 with photo #11, we see that the classifier can correctly distinguish between home and non-home as long as the background occupies roughly half of the photo.

In Figure 6, we show the prediction accuracy of four methods with increasing distance error tolerance. Clearly, our fusion predictor outperforms any other baseline methods with evident increase in prediction accuracy at every resolution level, from 70 meters to 1 km. Numerically, for the 70-meter distance tolerance, the relative improvement for the fusion predictor is 6% compared to photo feature alone, 12% compared to temporal feature alone and 16% compared to the baseline. With distance error tolerance equal to 1 km, the fusion predictor achieves a high accuracy at 79%. To put this in perspective, the New York City area covers a land area of 790 km^2 and the San Francisco area covers 121 km^2.

To further illustrate the reliable prediction performance of the fusion home predictor, Figure 7 shows two representative user examples, where example (a) is an incorrect home prediction of a user from the greater New York Area and example (b) is a correct home prediction of a user from the Bay Area. In example (a), we see that both photo #1 and #2 are taken indoors. However, human eyes can tell from the light screen and the empty room that photo #2 is much more likely to be taken at a photo studio rather than at home, while the computational approach cannot identify such subtlety. Also, we noticed that user (a) took a fair amount of various portrait photos at location #2, which further implies that location #2 is his or her working place. Due to these reasons, the fusion home predictor understandably assigned a high probability of being home to location #2.

The positive performance of fusion predictor indicates that the visual and the temporal feature provides complementary information to each other. For example, restaurant is a type of location where temporal feature can help the visual content. A photo of someone eating at restaurant is likely to be classified as eating at home, but the time and the frequency people dining out is different from that people stay at home. Thus, the unique temporal features can help the classifier distinguish between a restaurant and someone’s home. On the other hand, offices is a typical example where visual feature can help the temporal feature. Since people spend a lot of time at work, sometimes even during the night, it is possible for a classifier to mistake an office with home

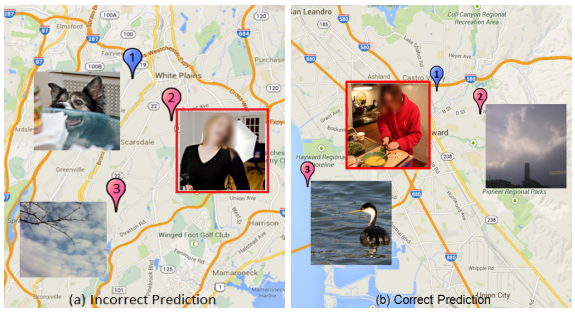

Figure 7: Two representative user examples showing the performance of our home predictor. For each user, the three pins represent the top 3 most frequently-visited locations, with home colored as blue and non-home locations colored as pink. The location marked in red box is predicted as home.

by using temporal feature alone. However, based on the visual content, the photo classifier can filter out offices to a certain extent.

In addition, the home classifier with photo feature alone outperforms the classifier with temporal feature for all distance error tolerances. It implies that the visual feature offers more reliable and definite clue to home location prediction.

Vacation Location Inference

The accurate home location prediction of Flickr user allows us to better understand and predict other points of interest. In this paper, we further propose a robust approach to predict a Flickr user's vacation locations based on the predicted home location and photo content.

Similar to home, a vacation location should also be user-specific since the same place might be a vacation spot to some people but not to others. From the spatial aspect, vacation locations should be away from home, say, at least 200 miles. For example, if a person living in Los Angeles went to Santa Monica beach, it should not be considered as a vacation since it is only about half an hour driving from the center of Los Angeles. However, if a person from New York checked in at Santa Monica beach, it is highly possible that he/she was going on a vacation. Therefore, we again used a pair *(user i, location j)* to differentiate a location checked in by different users.

Since a vacation spot can cover a large area from several square kilometers as a beach to tens of square kilometers in a national park, we retained two decimal places of each location's latitude and longitude values and clustered all photos by their geographic location. The error distance between the original geo-coordinate and the rounded geo-coordinate varies with the latitude and is bounded by 1.67 km.

To predict vacation locations, we kept the users who took at least 100 photos outside their home city and are left with 404 such users. These users have taken 423047 photos worldwide and resulted in about 31000 unique *(user, location)* samples. Ground truth is obtained by manually checking the photo collection at each location and thus we also filtered out those *(user, location)* samples with less than 20 photos since we are unable to determine the category of a location with less than 20 photos. This process resulted in 4142 unique *(user, location)* samples and manual checking gives us 900 vacation locations and 3242 non-vacation locations.

Spatiotemporal Feature

Similar to home location, vacation locations should share some temporal characteristics that is important to vacation inference.

The Flickr dataset shows an imbalanced distribution of vacation trips across the year: August, July, May and April are the top four most popular months for vacation while December and February are the off-seasons for vacation. Note that only 5 percent of the vacations are in December and this phenomena is consistent with the previous discover that people tend to stay at home during December.

Another important temporal feature is that people should go to a place for vacation only once or twice during a year and are expected to take a large volume of photos within a few days. So we again used the number of active months as a feature and discovered that 73% of the vacation locations are active for less than or equal to two months. Since we expect a large volume of photos to be taken during a vacation trip, we define two metrics to measure the *efficiency* of the photo-taking activity at each location. For each *(user i,location j)* sample, we define its *raw efficiency* as:

$$\frac{\text{\# of photos taken at at } \textit{(user i, location j)}}{\text{\# of active days at } \textit{(user i, location j)}} \tag{4}$$

Since users show different order of magnitude when taking photos, we define user i's *average efficiency* as:

$$\frac{\text{\# of photos taken by user } i}{\text{\# of active days of user } i \text{ across the year}} \tag{5}$$

and divided the raw efficiency by its corresponding user's average efficiency to get a normalized measure of the photo-taking efficiency for each *(user i, location j)* sample. Altogether we have extracted 15 temporal features including the check-in rate, # of active months, monthly rate and normalized efficiency.

Besides exploiting temporal features, we also want to filter out locations that are very close to home by using the spatial feature. For each *(user i, location j)* sample, we applied a sigmoid function with the origin at 1000 km to normalize the distance between location j and user i's predicted home location obtained from the previous experiment into the interval [0,1]. This normalized distance was fused with the set of temporal features mentioned above to make up a 16-dimensional spatiotemporal feature vector as our spatiotemporal baseline. Using 10-fold cross-validation and a Bayesian Network classifier over the 4142 samples, we find that the classifier has 0.468 precision and 0.468 recall, and 0.781 AUC (Area under curve) as depicted in Figure 9, indicating that the overall quality of the spatiotemporal baseline is decent.

One shortcoming of the spatiotemporal baseline is that it cannot identify a vacation location if the user did not take

a sufficient amount of photos during that vacation trip. On the other hand, misclassification of non-vacation events as vacations may occur for situations such as: international students going back to home country during school break, commencement, birthday and other occasions where a burst of photos will be taken, as well as business trips and academic visits. Therefore, to determine vacation locations more accurately, we extract the visual content from Flickr users' photo collections as a complementary clue to vacation inference.

Visual Feature

The photo collection at a vacation location should represent some natural or city scenes such as forest scenes, beach scenes and building scenes. In order to recognize the photo-taking scenes, we manually selected 35 categories of vacation-likely photos from the SUN Database (Xiao et al. 2010) as the training dataset, with some examples shown in Figure 8. We trained another CNN on this independent image dataset and generated a 35-dimensional score vector for each photo representing the probability of this photo belonging to the corresponding vacation category. With the visual feature as the second baseline, the visual feature predictor achieved 0.787 and the precision-recall curve is mostly above the one for the spatiotemporal baseline, as shown in Figure 9 .

Figure 8: Sample vacation photos representing ocean, hills, basilica, bridge, camping and harbor selected from SUN Database.

Visual feature performs well when people took a significant amount of scenic photos during the vacation. However, false negatives may occur for users who only take photo of indoor scenes (e.g. food, shows, and museums) during the vacation, and false positives may occur for parks or lakes near home. Therefore, to further improve the robustness of our vacation-predictor, we fused the spatiotemporal features with the visual feature to obtain a fused vacation predictor.

By fusing spatiotemporal and visual features, we obtain the red precision-recall curve shown in Figure 9 and AUC is now up to 0.854. The highlighted round points, which are the intersections between the precision-recall curve and the 45 degree line from the origin, show precision and recall both equal to 0.468, 0.507 and 0.594 for the spatiotemporal baseline, visual baseline and the fused vacation-predictor, respectively. The highlighted triangles are the points where the F1-measures are maximized, at 0.514, 0.524 and 0.609 for spatiotemporal, visual and the fused vacation predictor, respectively. These results indicate that the fused predictor outperforms the two baselines with respect to different metrics.

To further illustrate the robustness of the fused vacation predictor, we show two user examples in Figure 10. The user of the left example lives in Los Angeles and checked in at Las Vegas (location #1) and the Levis&Clark National Forest (location #2) in Montana state. Location #2 is correctly classified as vacation but location #1 is misclassified as non-vacation. Most photos at location #1 were taken at the famous St. Mark's Square in Las Vegas and it is clearly a vacation trip. However, since the photo are taken indoors and the user only took a small amount of photos there, the vacation predictor misclassified it as non-vacation. The example on the right shows a user living in New York and both two locations he/she visited are correctly classified. Location #2 is the Central Park in New York and it is classified as non-vacation since it is near the user' predicted home. Location #1 represents some mountain views and is classified as vacation.

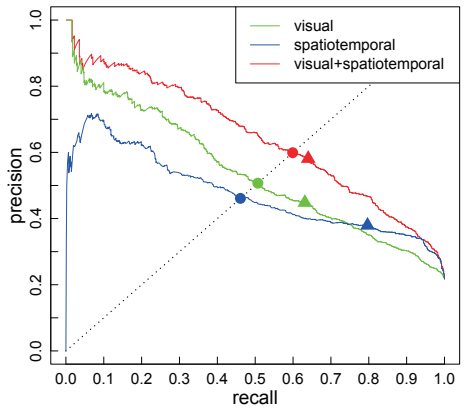

Figure 9: The performance of the baselines and the fused vacation predictor.

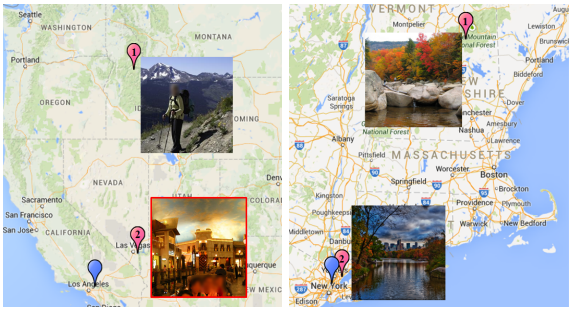

Figure 10: Two user examples showing the performance of the fused vacation predictor. For each user, the blue pin represents his/her predicted home location and the pink pins represent two of the user's visited locations shown with a representative photo taken at that location. The location marked in red is misclassified.

Conclusion and Future Work

In this paper, we present a novel multi-source approach to predicting Flickr users' POI locations with high precision and accuracy. The home predictor achieves an accuracy of 71% with a 70.7 meter error distance and the vacation predictor shows 0.594 precision and 0.594 recall. To accomplish this, we extract various features from a user's geotagged photos posted online. We employ a deep learning engine to semantically label photos to explore the visual content of real-life photos. By manually checking the results, we are convinced that our photo classifier based on CNN performs at a satisfactory precision in distinguishing real-life photos (Figure 5), compared with an SVM based scene recognition classifier (Figure 4). In addition to the visual content, we also take advantage of the temporal and spatial features of one's mobile trace as indicated by the photo geotags, such as the visiting rate of a location and the temporal regularity of a user's movement. Facilitated by the synergy of these features, our predictors for both home and vacation locations achieve remarkable overall performance.

In the future, we will expand the POI location category to include other significant locations such as work places. Moreover, based on the predicted home and vacation experience of Flickr users, we can build a vacation recommendation system that optimizes both location and time of the year. We also plan to improve our home detection method by adding richer spatio-temporal features such as the distance between the consecutive locations visited by people.

Acknowledgements

This work was generously supported in part by Google Faculty Award, Xerox Foundation, TCL Research America, the Intel Science & Technology Center for Pervasive Computing(ISTC-PC), NSF Award 1319378, NIH Award 5R01GM108337-02 and Adobe Research.

References

Cao, L.; Yu, J.; Luo, J.; and Huang, T. S. 2009. Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression. In *ACM MM*, 125–134. ACM.

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, 759–768. ACM.

Cho, E.; Myers, S. A.; and Leskovec, J. 2011. Friendship and mobility: user movement in location-based social networks. In *SIGKDD*, 1082–1090. ACM.

Gao, H.; Tang, J.; Hu, X.; and Liu, H. 2013. Modeling temporal effects of human mobile behavior on location-based social networks. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, 1673–1678. ACM.

Hays, J., and Efros, A. A. 2008. Im2gps: estimating geographic information from a single image. In *CVPR*, 1–8. IEEE.

Hoh, B.; Gruteser, M.; Xiong, H.; and Alrabady, A. 2006. Enhancing security and privacy in traffic-monitoring systems. *Pervasive Computing, IEEE* 5(4):38–46.

Jeffries, A. 2013. The man behind flickr on making the service awesome again.

Jia, Y. 2013. Caffe: An open source convolutional architecture for fast feature embedding. *h ttp://caffe. berkeleyvision. org*.

Joshi, D., and Luo, J. 2008. Inferring generic activities and events from image content and bags of geo-tags. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, 37–46. ACM.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, 1097–1105.

Krumm, J., and Rouhana, D. 2013. Placer: semantic place labels from diary data. In *UbiComp*, 163–172. ACM.

Liao, L.; Fox, D.; and Kautz, H. 2005a. Location-based activity recognition. In *NIPS*.

Liao, L.; Fox, D.; and Kautz, H. 2005b. Location-based activity recognition using relational markov networks. In *IJCAI*.

Luo, J.; Yu, J.; Joshi, D.; and Hao, W. 2008. Event recognition: viewing the world with a third eye. In *ACM MM*, 1071–1080. ACM.

Pontes, T.; Magno, G.; Vasconcelos, M.; Gupta, A.; Almeida, J.; Kumaraguru, P.; and Almeida, V. 2012a. Beware of what you share: Inferring home location in social networks. In *ICDMW*, 571–578. IEEE.

Pontes, T.; Vasconcelos, M.; Almeida, J.; Kumaraguru, P.; and Almeida, V. 2012b. We know where you live: privacy characterization of foursquare behavior. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 898–905. ACM.

Ross, G.; Jeff, D.; Trevor, D.; and Jitendra, M. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*.

Witten, I. H., and Frank, E. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR 2010*, 3485–3492. IEEE.

Ye, M.; Janowicz, K.; Mülligann, C.; and Lee, W.-C. 2011. What you are is when you are: the temporal dimension of feature types in location-based social networks. In *SIGSPATIAL*, 102–111. ACM.

Yuan, J.; Luo, J.; and Wu, Y. 2010. Mining compositional features from gps and visual cues for event recognition in photo collections. *Multimedia, IEEE Transactions on* 12(7):705–716.

Zheng, D.; Hu, T.; You, Q.; Kautz, H.; and Luo, J. 2014. Inferring home location from user's photo collections based on visual content and mobility patterns. In *Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia*, GeoMM '14. ACM.