Inferring Home Location from User's Photo Collections based on Visual Content and Mobility Patterns

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

ABSTRACT

Precise home location detection has been actively studied in the past few years. It is indispensable in the researching fields such as personalized marketing and disease propagation. Since the last few decades, the rapid growth of geotagged multimedia database from online social networks provides a valuable opportunity to predict people's home location 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 developed a reliable photo classifier based on the Convolutional Neutral Networks to classify photos as either home or non-home. We then proposed a novel approach to home location prediction by fusing together the visual content of web images and the spatiotemporal features of people's mobility pattern. Using a linear SVM classifier, we showed that the robust fusion of visual and temporal feature achieves significant accuracy improvement over each of the features alone.

Categories and Subject Descriptors

I.5.4. [**Pattern Recognition: Applications.**]: Miscellaneous, Data Mining

General Terms

Algorithms, Experimentation

Keywords

home location, mobility pattern, home picture recognition

1. INTRODUCTION

Precise home location is increasingly important in various researching fields. Home location information is indispensable in the study of geographic mobility since home is such a crucial node in people's activity trace. In urban planning, knowing location-based behavior can help build more personalized design of urban environment, including the transportation networks and pollution management. Besides these, other important researching fields such as disease propagation and outbreak modeling all require researchers knowing where people live.

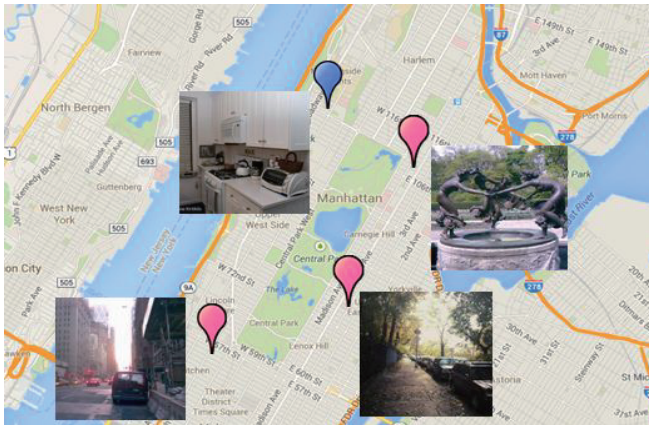

Figure 1: Visualization of a Flickr user's activity trace in New York City. The four pins represent the top 4 most frequently-visited locations, with home colored as blue and non-home locations colored as pink. Each pin is shown with a photo taken at that location.

Existing methods which can precisely detect home location are based on surveys, GPS data or cellular telephone records [11, 7, 4]. However, the process of obtaining such continuous data are often time and labor consuming. 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 [1], these information are not adaptable for follow-up investigations since we cannot combine the user information 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 [3] or check-ins [14]. One of the main existing issues is that these methods either suffer from coarse granularity, at city [13] or state level, or result in a low accuracy, at around 50% [3].

GeoMM'14, November 7, 2014, Orlando, FL, USA.

http://dx.doi.org/10.1145/2661118.2661123.

In this paper, we addressed this home prediction problem by analyzing photos mined from Flickr. As a popular image-hosting online community, Flickr has more than 3.5 million new images uploaded per day [8]. 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.

The result showed that the visual content of images can provide valuable clues complementary to the metadata captured with photos and can be used to improve home location prediction performance. We believe this is the first time home location is predicted at such a fine-grained scale by mining informative visual features from images collected from social networks.

The contributions made in this study are thus twofold. First, we developed a reliable classifier by the Convolutional Neutral Networks [10], which can recognize the photo-taking scene as either "home" or "non-home" of real-life photos. Second, we fused the visual content of web images with the spatiotemporal features of a user's online photo-sharing activity to construct a robust multi-source home predictor, where each of the two features contributes to the improvement in home location. The precision to which we can locate a person allows various location-related research in greater depth and with higher accuracy.

2. RELATED WORK

Locations such as home, working places and restaurants are important in understanding human mobility pattern and automatically predicting human's future activity. In [11], Krumm et al. 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 spend doing various activities[1]. Krumm et al. used demographic and temporal features of people's activity to infer a place's label and the result showed that home location can be predicted with a high accuracy at 92%.

As people spend more time online, social networks enable an alternative approach to semantically label geographic locations. Cheng et al.[3] 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 51% accuracy.

On the other hand, our work is also closely related to the study of semantic and geographic annotation of web images[12, 19, 6, 2]. As photographic equipments 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. [6] estimated the geographic location of an image based solely on its image content, including color, texture and line features. Based on a series of geotagged photos [19], Yuan et al. detected the associated event by fusing visual content and the spatiotemporal traces of geographic coordinates. The 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 [12].

3. DATA

In this section, we describe how we obtained the ground truth (users' actual home locations) and built the dataset used to train and evaluate the machine learning model.

Instead of using user profiles, we used the geotags of a user's uploaded photos to locate his or her actual home. We selected a set of tags containing "in home", "in kitchen", "in bed", "family time" and their variants, and refer to them as home-related tags. Note that we have manually checked the photos tagged with these home-related words to make sure that most of the returned photos are highly related with home. Using the Flickr Search API, we collected all photos with home-related tags in the Bay Area and the greater New York City Area. We manually filtered out the photos that are not taken at home and then clustered these photos by users. Altogether, we have mined 2167 photos taken at home by 192 unique users.

For each user i, we recorded a sequence $t_i = (t_{i1}, t_{i2}, ...)$, where t_{ij} represents the time point of user i's jth photo taken at home. In consideration of home moving, we queried Flickr for all public photos posted by these 192 users in a one-year time length, which is obtained by adding and subtracting half-a-year from the median time point in sequence t_i. Each photo is associated with a geographic tag accurate to the street-level, which is represented by a pair of longitude and latitude coordinates. Altogether, we have collected 47793 photos taken by 192 users in a one-year time length. Then we divided the Bay Area and the greater NYC Area 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 larger than 70.7 meters.

4. METHODS

This section presents the methods we used to quantify a Flickr user's online photo-sharing activity and predict his or her home location. For each user, an uploaded photo with geographic tag is referred to as one upload event and is considered as a visit to that geographic location. In the Flickr dataset, there exist some 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 tuple (i, j) as an ID to represent a location j being visited by user i. Altogether, we have recorded 8675 unique (*user*, *location*) IDs by 192 users.

4.1 Temporal Features

According to previous work [14], 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 model as the baseline and refer to it as the most-visited method. Built on this baseline, we then mine a large collection of temporal features for each unique (*user*, *location*) ID. As validated in previous work [18, 5], human mobile behavior displays strong temporal cyclic patterns and this temporal regularity can help

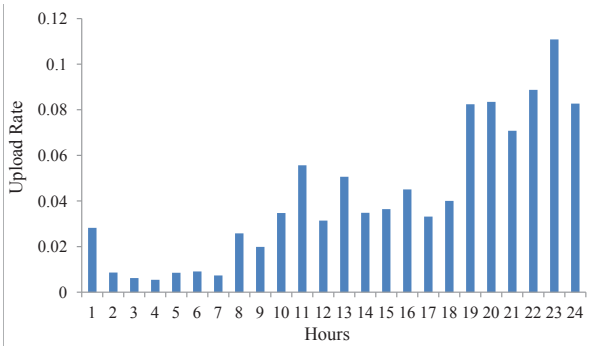

Figure 2: Upload rate distribution at home on an hourly basis. Y-axis represents the percentage of the # of "home photo" uploaded during a specific hour.

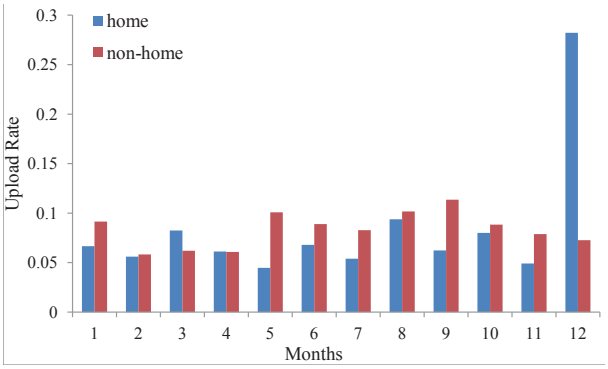

Figure 3: Comparison of the # of uploads at home/non-home locations on a monthly basis. Y-axis represents the percentage of the # of "home photo"/"non-home photo" uploaded during a specific month.

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 feature by adding visual content feature to our previous collection of temporal features and compare the performance of home prediction.

Similar to previous work [18, 5], the Flickr data set shows strong evidence of yearly patterns (months across a year) and daily patterns (hours across a day) of a Flickr user's online photo-sharing activity. In Fig. 2, we see that the number of photos uploaded at home roughly follows an ascending trends from 3 am to midnight. The few hours before midnight are the most active time slot at home while the distribution decays rapidly after midnight. It reveals that if a user is still active after midnight then he or she is more likely to be somewhere else such as night clubs or parties rather than at home. Fig. 3 demonstrates a significant difference between the number of uploads at home and non-home locations on a monthly basis. December stands out from all the months in the sense that the number of uploads at home in December is significantly higher than that in other months. Numerically, among all photos uploaded at home, the number of photos uploaded in December accounts for nearly 30% of the total. Note that this phenomena is specific for home

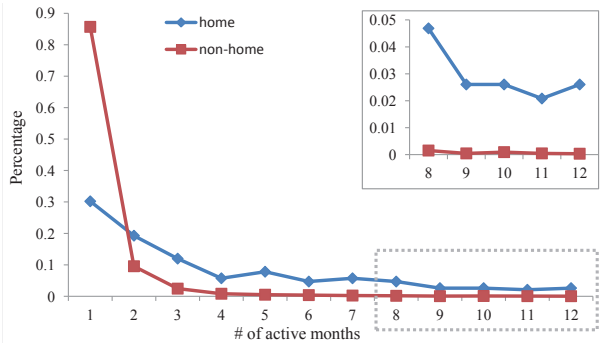

Figure 4: Comparison of the # of active months at home/non-home locations during a year. Y-axis represents the percentage of home/non-home that are active for a specific # of months. The plot on top right corner is a magnification of the part in the gray dotted area.

since the number of uploads 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 the Christmas and share plenty of photos during that time.

Another interesting observation is that the number of active months (a month is referred to as an "active month" if the user uploads at least one image during that month) at home are universally larger than that at non-home locations, as home photos can be taken at any time of the week, and any month of the year. As shown in Fig. 4, more than 50% of the home locations are active for at least three months while less than 10% of the non-home locations are active for more than 2 months. This distribution reveals that although people may upload a massive amount of photos during certain events such as traveling, such events would only happen once or twice during a year.

Clearly, there exists a high correlation between time and the number of uploads in the Flickr dataset. Therefore, we extract a large collection of temporal features to represent each unique (*user*, *location*) ID. 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, January upload rate is given by:

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

Altogether, for each (*user*, *location*) ID, we extracted 40 temporal features, including monthly upload rate, hourly upload rate, weekday upload rate, weekends upload rate, # of active hours out of a day and # of active months out of a year.

4.2 Visual Features

Different from tags and descriptions of online photos, which are usually not available or informative enough, visual content is mandatory for each photo. As an inherent feature, visual content provides us fundamental insight of 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 hidden in photos' visu-

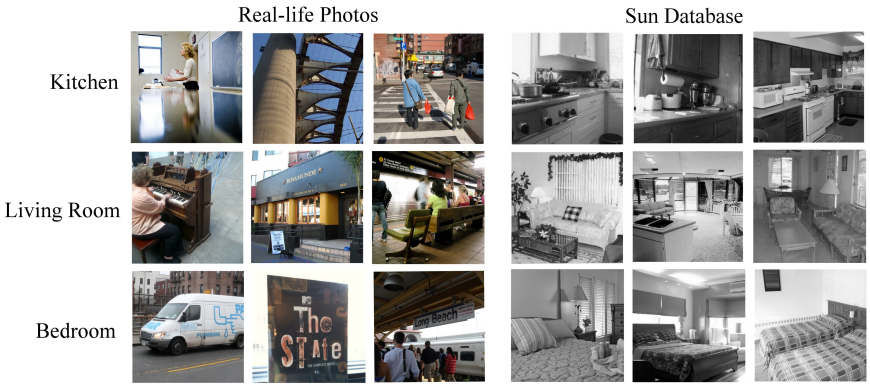

Figure 5: Examples of real-life and sun database photos classified as "kitchen", "living room" and "bedroom" by using HOG 2×2 features.

al content, we trained a classifier to distinguish "home-like" photos from the others.

Scenes recognition approaches can be employed to extract the visual content of pictures. In [17], HOG2×2 features were used to classify photos into 397 categories (e.g. living room, kitchen) and it achieved a higher accuracy than other single feature based methods. To distinguish "home-like" photos 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 presence of real-life photos make it extremely challenging for classification. In Fig. 5, we show the classification result of HOG2×2+SVM. The classification produces desirable results on the Sun database, but it performed poorly when we applied it on real-life photos.

Inspired by some recent successes [15], 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 a "home photo" or a "non-home photo". For each (*user*, *location*) ID, we define "home photo" rate as:

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

and use it as the visual content feature. As described in [10], we extract a 4096-dimensional feature vector for each photo by using the Caffe [9] implementation of the Convolutional Neutral Networks. We pre-trained the CNN on an image dataset with manual labels and fine-tuned the network by iteratively feeding back false positive and false negative images to the training set.

With the ground truth and the features mentioned above, we trained an SVM-based metaclassifier using the Weka toolkit [16] over the set of (*user*, *location*) IDs. 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 most-visited baseline method. In our experiment, two-fold cross-validation is used to test the robustness of our methods.

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

A few representative examples of photos are presented in Fig. 6 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 sofas, tables and stoves (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 #6) and a discarded TV on the street (photo #4) 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 approach. The "non-home photo" examples reveal that the photo classifier can accurately identify outdoor photos even for a portrait-oriented photo. Comparing photo #3 with photo #11, we see that the classifier can correctly distinguish between home and non-home as long as the background covers roughly half of the photo.

In Fig. 7, 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

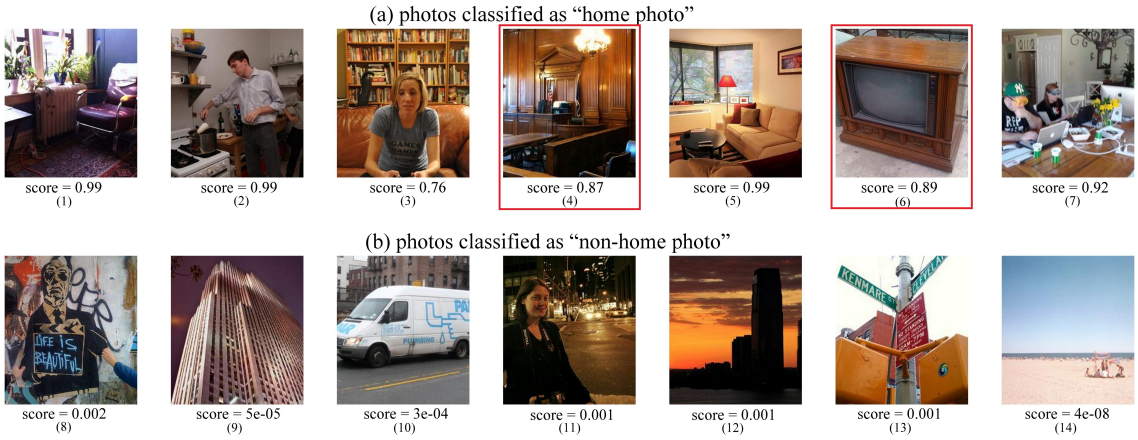

Figure 6: Examples of photos classified by the trained deep networks as (a) "home photo" and (b) "non-home photo". Photos marked in red boxes are misclassified.

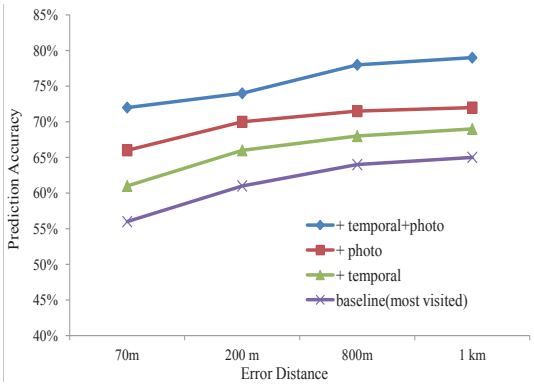

Figure 7: The performance of the baseline and the fusion home predictors. The plot shows the prediction accuracy with increasing distance error tolerance(70 meters to 1000 meters).

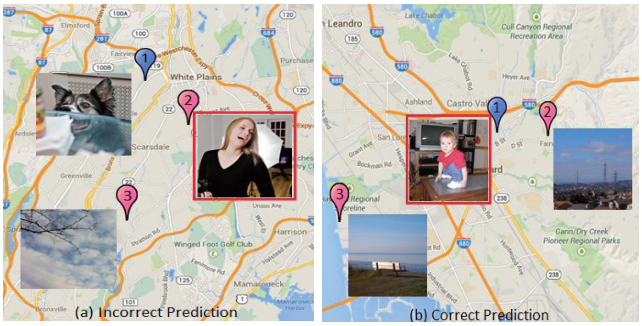

Figure 8: Two representative user examples showing the performance of home predictor. For each user, the three pins represent the top 3 most frequently-visited locations, with home colored as blue and non-home locations colored as pink. The location marked in red box is predicted as home.

this in context, the New York City covers a land area of 790 km^2 and the San Francisco city covers 121 km^2.

To further illustrate the reliable prediction performance of the fusion home predictor, Fig. 8 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 film studio rather than at home, while the computational approach cannot identify such subtle details. 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 reasonably 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 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.

6. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel multi-source approach to predicting Flickr users' homes within 70 meters with an accuracy of 72%. To achieve this, we extract various features

from a user's geotagged photos posted online. We employ a deep learning engine to reliably label photos as "home" or "non-home" 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 (e.g. Fig. 6), compared with single-feature based scene recognition classifier (e.g. Fig. 5). In addition to the visual content, we also take advantage of temporal and spatial features of one's mobile trace as indicated by the photo geotags, such as the visit rate of a location and the temporal regularity of a user's movement. Facilitated by the complementary effect of these features, our predictor achieves a promising overall performance. Evaluated on the ground truth, our method performs better than any other baseline methods at every resolution level. In particular, with distance error tolerance equal to 1 km, the fusion predictor achieves a high accuracy at 79%.

In the future, we will extend our method to detect other important places such as schools, working places and vacation spots. It is also interesting to use our method to study people's mobility patterns, life styles, and so on. We can also improve our home detection method by adding richer spatio-temporal features such as the distance between the locations visited by people.

7. REFERENCES

[1] American time use survey user's guide. http://www.bls.gov/tus/atususersguide.pdf.

[2] L. Cao, J. Yu, J. Luo, and T. S. Huang. Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 125–134. ACM, 2009.

[3] Z. Cheng, J. Caverlee, and K. Lee. 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*, pages 759–768. ACM, 2010.

[4] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[5] H. Gao, J. Tang, X. Hu, and H. Liu. 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*, pages 1673–1678. ACM, 2013.

[6] J. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

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

[8] A. Jeffries. The man behind flickr on making the service Śawesome again.Š, 2013.

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

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[11] J. Krumm and D. Rouhana. Placer: semantic place labels from diary data. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 163–172. ACM, 2013.

[12] J. Luo, J. Yu, D. Joshi, and W. Hao. Event recognition: viewing the world with a third eye. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 1071–1080. ACM, 2008.

[13] T. Pontes, G. Magno, M. Vasconcelos, A. Gupta, J. Almeida, P. Kumaraguru, and V. Almeida. Beware of what you share: Inferring home location in social networks. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, pages 571–578. IEEE, 2012.

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

[15] G. Ross, D. Jeff, D. Trevor, and M. Jitendra. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

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

[17] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010.

[18] M. Ye, K. Janowicz, C. Mülligann, and W.-C. Lee. What you are is when you are: the temporal dimension of feature types in location-based social networks. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 102–111. ACM, 2011.

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