Venue Prediction for Social Images by Exploiting Rich Temporal Patterns in LBSNs

Jingyuan Chen[1(✉)], Xiangnan He[1], Xuemeng Song[2], Hanwang Zhang[3], Liqiang Nie[2], and Tat-Seng Chua[1]

[1] School of Computing, National University of Singapore, Singapore, Singapore
jingyuanchen91@gmail.com, xiangnanhe@gmail.com, dcscts@nus.edu.sg
[2] School of Computer Science and Technology, ShanDong University, Jinan, China
sxmustc@gmail.com, nieliqiang@gmail.com
[3] Department of Computer Science, Columbia University, New York, USA
hanwangzhang@gmail.com

Abstract. Location (or equivalently, "venue") is a crucial facet of user generated images in social media (*aka.* social images) to describe the events of people's daily lives. While many existing works focus on predicting the **venue category** based on image content, we tackle the grand challenge of predicting the **specific venue** of a social image. Simply using the visual content of a social image is insufficient for this purpose due its high diversity. In this work, we leverage users' check-in histories in location-based social networks (LBSNs), which contain rich temporal movement patterns, to complement the limitations of using visual signals alone. In particular, we explore the transition patterns on successive check-ins and periodical patterns on venue categories from users' check-in behaviors in Foursquare. For example, users tend to check-in to cinemas nearby after having meals at a restaurant (transition patterns), and frequently check-in to churches on every Sunday morning (periodical patterns). To incorporate such rich temporal patterns into the venue prediction process, we propose a generic embedding model that fuses the visual signal from image content and various temporal signal from LBSN check-in histories. We conduct extensive experiments on Instagram social images, demonstrating that by properly leveraging the temporal patterns latent in Foursquare check-ins, we can significantly boost the accuracy of venue prediction.

Keywords: Venue prediction · Transition pattern · Temporal pattern

1 Introduction

The unprecedented growth of smart mobile devices allows people to easily take pictures of their life events and post them on online social networks. As a result,

This research is part of NExT++ project, supported by the National Research Foundation, Prime Ministers Office, Singapore under its IRC@Singapore Funding Initiative and National Natural Science Foundation of China under Grant No.: 61702300.

K. Schoeffmann et al. (Eds.): MMM 2018, Part II, LNCS 10705, pp. 327–339, 2018.
https://doi.org/10.1007/978-3-319-73600-6_28

the current social Web is experiencing a tremendous volume of user generated images, which we term as social images [6]. According to a recent study by Chen *et al.* [7], over 45% of tweets are associated with images in Weibo — the largest micro-blog service in China. To facilitate understanding and use of such social images, it is crucial to address the fundamental problem of "where did it happen".

Fig. 1. Examples of social images with venue tags to show the challenge of venue prediction based on visual content only.

Several existing works have explored the prediction of venue category for social images [22] and micro-videos [21], such as to predict whether a photo is taken at a *restaurant* or a *theme park*. In this work, we move one step further, to predict the **specific venue**[1] of a social image, which permits us to infer a user's footprint more accurately so as to provide better location-based services. Specifically, we aim to predict the exact venue where the user was taking the photo, such as the *Los Tacos No. 1 restaurant* or *Universal Studios Singapore*, rather than the general categories of *restaurant* or *theme park*. Nevertheless, existing approaches rely solely on images' visual contents [21,22] for venue category prediction, which is far from being sufficient to predict the specific venue accurately due to the high diversity of social images. Figure 1 gives an illustrative example. The first row shows that simply relying on visual content of the image itself, it is difficult to distinguish whether the image is taken in the Universal Studio of *Singapore*, *Osaka* or *Hollywood*. The second row shows that images taken at the same venue can be very different visually.

To alleviate the difficulty, an intuitive idea is to utilize a user's historical locations to restrict the venue prediction candidates and discover possible

[1] In this work, we use point-of-interest (POI), venue, and location interchangeably, which all refer to a specific venue.

movement patterns. As an example, the visual content may show that the image was taken in a *McDonald's restaurant*, while the recent movement history of the user can help to identify which specific store of *McDonald's*. However, the high sparsity of venue-tagging activities of social network sites (*e.g.,* Twitter and Instagram) makes it challenging to implement the idea based on the data from one site only. Fortunately, the emerging of location-based social networks (LBSNs) provides an excellent alternative source of data to tackle the problem. For example, users mainly use Foursquare[2] to check-in to POIs, leaving us with valuable spatial-temporal trails of users' historical movements. In this work, we explore the possibilities of mining check-in histories of LBSNs to address the problem of predicting the exact venues of images in a social media site with sparse check-in histories.

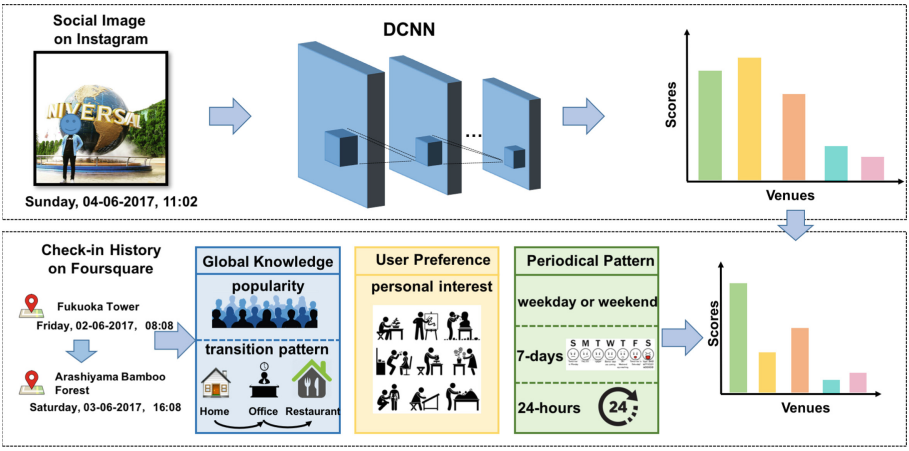

Fig. 2. Overview of our proposed framework. Given a social image and the user's check-in history, our framework predicts the specific venue of the image by inferring a probabilistic distribution. (a) We train a deep convolutional neural network to learn venue information (both category and specific venue) from visual content. (b) We mine various signals from the check-in histories of a LBSN to complement the visual signals.

Using Foursquare as a check-in–rich LBSN for our case study, we first conducted comprehensive statistical analysis to understand users' movement behaviors. Our analysis reveals some promising patterns that support our premise of using LBSN check-in histories for venue prediction. For example, we found that people typically move within a bounded region and seldom travel long distances within a short time; and moreover, successive check-ins usually exhibit certain correlations and strong periodical patterns. Guided by these phenomenon, we developed an end-to-end probabilistic solution for social image venue prediction (Sect. 3). Figure 2 shows an overview of our proposed framework. Specifically, our

[2] https://foursquare.com.

solution unifies the inference of venue category and specific venue, and both components carefully fuse the visual signal from image content and various temporal signal from LBSN check-in histories. We evaluate our solution on social images of Instagram (Sect. 4). Extensive experiments demonstrate that by exploiting the rich temporal patterns in Foursquare check-ins, we can significantly enhance the venue prediction accuracy by 5.7% on average.

2 Related Work

Image venue prediction, also called image geotagging, aims to identify the venue, landmark or location that an image refers to from a set of candidates [1,18,19]. Most methods extract a rich set of visual features from images and leverage the visual features to train either shallow or deep models to estimate the venues of the given images. As reported in [12], the landmark identification [3] and scene classification [2] of images are the key factors to recognize the locations. In addition to visual content, Crandall et al. [9] combine both visual and textual information to map photos on Flickr into different venue categories based on landmarks. These approaches are based on the observation that there exists strong correlation between the content of images with certain venue categories. However, most of these methods ignore the high diversity of the content of social images. For example, visually similar images can be taken at different venues, while images taken at the same places can have different visual appearance. In such cases, utilizing only the content of images is far from being sufficient to predict the specific venue accurately. Our work differs from these studies by utilizing a user's historical movement behaviors on LBSN to discover users' possible movement patterns and adjust the prediction accordingly.

3 Proposed Method

In this section, we begin by formulating the venue prediction problem, followed by elaborating the design of solution components one by one. In terms of techniques, we wish to develop models that are expressive enough to capture the various relevant signals and temporal dynamics, while at the same time can generalize well with a controllable number of parameters. To achieve these design goals, we resort to embedding-based models, which encode various features and patterns in the latent space.

3.1 Problem Formulation

Let $\mathcal{S}_u = \{s_u^1, s_u^2, \cdots, s_u^{n_u}\}$ be the historical check-in sequence for user u, where n_u denotes the number of u's historical check-ins; if a user has check-in behaviors on multiple social networks, we can merge them together to form $\mathcal{S}_u$. Each check-in entry s_u^n is represented as $s_u^n = (l_u^n, i_u^n, t_u^n)$, meaning the $l$ocation, $i$mage, and $t$ime of the check-in. Note that i_u^n is an optional field for s_u^n, as not all check-ins

are associated with images. Each location l_u^n corresponds to a venue category c_u^n. Given a set of images with venue categories, we train a ResNet to predict the venue category of the image. In this work, we aim to solve the problem of predicting the specific venue $l_u^{n_u+1}$ of the next check-in $s_u^{n_u+1}$, given its check-in image $i_u^{n_u+1}$, timestamp $t_u^{n_u+1}$, and the user's check-in history. From a probabilistic point of view, the task can be addressed by inferring the probability that user u would visit $l_u^{n_u+1}$ at time point $t_u^{n_u+1}$:

$$p(l_u^{n_u+1}|\mathcal{S}_u, i_u^{n_u+1}). \tag{1}$$

Apparently, it is not advisable to predict the specific venue directly. On one hand, it is more challenging to predict the specific venue than the venue category solely based on the visual content of social images due to their high diversity (example see Fig. 1). On the other hand, it has been found that there exists a strong correlation between the categories of two successive check-ins for a user in a short period according to our pilot study. As such, we tackle the problem in two steps. First, for each social image, we predict its venue category by taking into account multiple aspects including the visual content, user's personal interests, global popularity of the category and the transition probability between successive check-ins. Second, given the predicted venue category, we predict the specific venue based on the similar aspects. Mathematically, we decompose Eq. (1) into two parts:

$$p(l_u^{n_u+1}|\mathcal{S}_u, i_u^{n_u+1}) = p(l_u^{n_u+1}|\mathcal{S}_u, i_u^{n_u+1}, c_u^{n_u+1}) \cdot p(c_u^{n_u+1}|\mathcal{S}_u, i_u^{n_u+1}). \tag{2}$$

In what follows, we first detail how to predict the venue category, and then discuss the venue prediction model. We pay special attention to temporal modeling, which is used in both the category and venue prediction.

3.2 Venue Category Prediction

Although there are several work on predicting venue category of multimedia posts on social networks, such as the micro-videos on Vine [5,21], these work are merely based on the multimedia (and textual) content while overlooking other relevant signals. One important signal is the sequential pattern between two successive check-in categories. For example, if a user just checked in at a *shopping mall*, then the user is more likely to visit a *restaurant* as compared to *office* or *school*. Such sequential patterns naturally motivate us to adopt the Markov chain [8] modeling. Secondly, venue categories usually show varying global popularity, for instance, users tend to check in at *office* more frequently than *hospital*. Apart from the global popularity, they may also express different personal interests in different venue categories. To tackle the venue category prediction problem in a comprehensive way, we summarize the key factors to be accounted for as follows:

- the global popularity of venue categories;
- the consistency between the user's personal interest and the given venue category;
- the correlation between the venue category of the last check-in and that of the current check-in;
- the matching of the visual content of social image and the given venue category.

Based on the first-order Markov chain property, the venue category probability conditioning on the full check-in sequence can be approximated as conditioning on the last check-in:

$$p(c_u^{n_u+1}|\mathcal{S}_u, i_u^{n_u+1}) = p(c_u^{n_u+1}|c_u^{n_u}, i_u^{n_u+1}). \tag{3}$$

By incorporating the proposed factors, we further parameterize the probability as:

$$p(c_u^{n_u+1}|c_u^{n_u}, i_u^{n_u+1}) = \alpha_u^{n_u+1} + \tilde{\mathbf{v}}_u \cdot \tilde{\mathbf{c}}_u^{n_u+1} + \tilde{\mathbf{c}}_u^{n_u} \cdot \tilde{\mathbf{c}}_u^{n_u+1} + \theta_1 p(c_u^{n_u+1}|i_u^{n_u+1}), \tag{4}$$

where $\alpha_u^{n_u+1}$ denotes the global popularity of category $c_u^{n_u+1}$; vectors $\tilde{\mathbf{v}}_u \in \mathbb{R}^{D_1}$ and $\tilde{\mathbf{c}}_u^{n_u} \in \mathbb{R}^{D_1}$ denote the embedding vector for user u and category $c_u^{n_u}$, respectively. The first inner product $\tilde{\mathbf{v}}_u \cdot \tilde{\mathbf{c}}_u^{n_u+1}$ encodes the personalized preference of user u on venue category $c_u^{n_u+1}$, and the second inner product $\tilde{\mathbf{c}}_u^{n_u} \cdot \tilde{\mathbf{c}}_u^{n_u+1}$ encodes the transition probability between the two successive check-ins. Note that we have intentionally chosen inner product to model the interaction between two entities, which although is simple, but shown to be very competitive compared to more complex neural network functions recently [16].

The last term $p(c_u^{n_u+1}|i_u^{n_u+1})$ denotes the probability that image $i_u^{n_u+1}$ belongs to category $c_u^{n_u+1}$, which is inferred from the visual content of the image. In our implementation, we employ the state-of-the-art deep convolutional neural network ResNet [13], pre-training it on ImageNet data and fine-tuning it to our venue category data (more details see Sect. 4.1). The hyper-parameter θ_1 should be non-negative that controls the weight of the visual signal. Note that we have further normalized the probabilities such that $\sum p(\cdot|c_u^{n_u}, i_u^{n_u+1}) = 1$, to make it a valid conditional probability in the strict sense.

3.3 Specific Venue Prediction

As have mentioned before, most of user movements are constrained to a relatively small geographical region in a short period. As such, a social image is more likely to be taken at the venues near the user's last check-in. We similarly adopt the first-order Markov chain assumption and model the venue probability as:

$$\begin{aligned} p(l_u^{n_u+1}|\mathcal{S}_u, i_u^{n_u+1}) &= p(l_u^{n_u+1}|s_u^{n_u}, i_u^{n_u+1}) \\ &= \beta_u^{n_u+1} + \hat{\mathbf{v}}_u \cdot \hat{\mathbf{l}}_u^{n_u+1} + \hat{\mathbf{l}}_u^{n_u} \cdot \hat{\mathbf{l}}_u^{n_u+1} + \theta_2 p(l_u^{n_u+1}|i_u^{n_u+1}), \end{aligned} \tag{5}$$

where $\beta_u^{n_u+1}$ represents the global popularity of venue $l_u^{n_u+1}$, vectors $\hat{\mathbf{v}}_u \in \mathbb{R}^{D_2}$ and $\hat{\mathbf{l}}_u^{n_u} \in \mathbb{R}^{D_2}$ denote the embedding vector for user u and venue $l_u^{n_u}$, respectively. Note that the specific venue model shares a similar formulation with the venue category model, as users' movement patterns on venue categories can be smoothly transferred to specific venues. Specifically, the two inner product terms encode the personalized user preference and transition pattern, respectively. The last term $p(l_u^{n_u+1}|i_u^{n_u+1})$ denotes the probability that image $i_u^{n_u+1}$ is taken at venue $l_u^{n_u+1}$ judging from the visual content. The hyper-parameter θ_2 should be non-negative that controls the weight of the visual signal. As it is computationally expensive and inefficient to train deep learning models over hundreds of thousands venues, we estimate the probability as the cosine similarity between location features and image features. Specifically, the image features are extracted by the ResNet (trained on venue categories), and the location features are the average of extracted features of images that are taken at the location.

3.4 Modeling Temporal Dynamics

We now present how to incorporate other temporal dynamics into our prediction model. We discuss the details in the context of venue category prediction model. As the encoding of temporal dynamics to specific venue model can be achieved in a similar way, we omit these details here to avoid repetition.

Periodical patterns
According to our statistics, users' visits to some venue categories, such as *concert hall*, *office*, *pub* and *restaurant*, are highly dependent on the time (*e.g.,* day of the week) and exhibit periodical patterns. For example, users typically visit *pub* more frequently during the night of weekends than weekdays. Moreover, within the span of a day, the category distribution also varies temporally. For example, users prefer to check in more at *train station* in the morning, while check in more at *cafe* in the afternoon. To capture the influence of temporal information, it is natural to extend the user embedding vectors $\tilde{\mathbf{v}}_u$ and $\hat{\mathbf{v}}_u$ to be time-dependent. Specifically, we introduce three time-related categorical variables to interpret a timestamp:

- t_1 indicates whether the image is uploaded on weekday or weekend, *e.g.,* 0 means weekday and 1 means weekend.
- t_2 indicates the day of the week that the image is uploaded, *e.g.,* 0 represents Monday, 1 represents Tuesday, 2 represents Wednesday, *etc.*
- t_3 indicates the time of the day the image is uploaded, where a day is divided into 24 h from 0 to 23.

Accordingly, we introduce three types of embedding vectors to capture the temporal dynamics in user representation as,

$$\tilde{\mathbf{v}}_u(t) = \tilde{\mathbf{v}}_u + \mathbf{E}_1(t_1) + \mathbf{E}_2(t_2) + \mathbf{E}_3(t_3), \tag{6}$$

where $\mathbf{E}_1 \in \mathbb{R}^{D_1 \times 2}$, $\mathbf{E}_2 \in \mathbb{R}^{D_1 \times 7}$ and $\mathbf{E}_3 \in \mathbb{R}^{D_1 \times 24}$ denote the embedding matrix for the three time indicators, which are weekday or weekend, day of the week and time of the day, respectively. The symbol $\mathbf{E}_i(t_i)$ returns the t_i-th column of $\mathbf{E}_i$, and similarly for other notations. Here the stationary user preference is modelled by $\tilde{\mathbf{v}}_u$, while the time-dependent periodical patterns are accounted by remaining parts.

Temporal weighting drift

In addition to the periodical patterns, we also consider the temporal dependence between two successive check-ins. As the time interval between two successive check-ins generally follows a Poisson distribution and may range from several minutes to days, it is natural to assume that smaller interval leads to higher dependence. To capture this effect, we introduce a time decay component similar to [15] to adjust the transition probability:

$$p(c_u^{n_u+1}|c_u^{n_u}, i_u^{n_u+1}) = \alpha_u^{n_u+1} + \tilde{\mathbf{v}}_u(t_u^{n_u+1}) \cdot \tilde{\mathbf{c}}_u^{n_u+1} + \delta(\Delta t)\tilde{\mathbf{c}}_u^{n_u} \cdot \tilde{\mathbf{c}}_u^{n_u+1} + \theta_1 p(c_u^{n_u+1}|i_u^{n_u+1}). \tag{7}$$

where $\Delta t = t_u^{n_u+1} - t_u^{n_u}$ represents the time interval between two successive check-ins and $\delta(\Delta t) = e^{-\lambda \Delta t}$ is the rate of the decay. Through this, the dependency on previous check-in can be gradually weakened as time goes by.

Table 1. Number of data records in NUS-MSS.

City	users#	ch-ins#	images#	venues#	venue cats#
London	2,860	63,273	11,661	5,857	394
Singapore	5,677	284,258	10,711	14,010	450
New York	5,122	189,152	23,925	14,891	485

4 Experiments

We first present the experimental settings, followed by studying the performance of our proposed solution in venue prediction and exploring the effectiveness of temporal check-in patterns. Lastly, we perform some micro-level analysis by showing some illustrative examples.

4.1 Preliminaries

In this work, all the experiments are conducted based on the NUS-MSS dataset [11], which provides a set of users' behaviors on multiple social networks. In particular, we utilize users' check-in sequences on Foursquare and social images on Instagram. Given a social image i posted by user u on Instagram at time point t, user u's last check-in venue is his/her lastest check-in on Foursquare before time point t. To ensure the quality of the dataset, we

retain only users with at least 3 check-ins, and venues that have been visited at least 3 times. Finally, we obtain a dataset consisting of 2860, 5677 and 5122 users, 63273, 284258 and 18912 check-ins, and 11661, 10711 and 23925 images in London, Singapore and New York respectively, as shown in Table 1.

Dataset Alignment. Due to that the POI tags provided by Instagram cannot be directly aligned with those on Foursquare, we first need to tackle the problem of venue alignment. For each venue of Instagram, we crawl its profile to obtain the name and location information (i.e. longitude and latitude), based on which we utilize the Foursquare venue/search api endpoint[3] to link each venue in Instagram to that in Foursquare.

Visual Feature Extraction. As certain objects should frequently appear in certain venue categories or venues, the multimedia content of social images plays an important role in venue prediction. To build the deep model based on the visual contents of social images as mentioned in Sect. 3.2 and feature extraction of specific venue matching as mentioned in Sect. 3.3, for each venue category, we collected extra 200 images from the venue profile on Foursquare. We then employ the ResNet-50 [13] model, which has been extensively studied in computer vision domain, as the architecture of deep model. This model is trained on ImageNet [10] and then fine-tuned on the venue category dataset we collected. Finally, we adopt the output after *softmax* as the prediction score for each venue category and *pool5* layer output as the features for specific venue matching. It is noted that the venue feature vector is generated by averaging all the image features belongs to that venue.

Evaluation Metrics. To evaluate the performance of venue prediction, we adopt the leave-one-out evaluation strategy, where for each user, we select the latest social image as the testing sample and the remaining data for training. Regarding the evaluation metrics, on one hand, we use the average top-1 and top-5 accuracies, which are the standard measures for the single-label task [4]. On the other hand, to assess the position of the hit, we adopt another widely used metric—Normalized Discounted Cumulative Gain (NDCG) [17]. Finally, we report the average score for all testing samples. It is worth mentioning that to save the time cost, for each testing sample, we randomly sample 200 negative samples as the venue candidates rather than the whole samples.

4.2 Model Comparison

Since our proposed model is derived from the matrix factorization method [20] with temporal patterns, we term it as **MFTP** for short. To evaluate the proposed method, we compare it with the following baselines:

– **VenuePop.** Venues are ranked by their popularity, which is measured by the number of check-ins. This is a non-personalized method to benchmark the prediction performance.

[3] https://developer.foursquare.com/docs/.

Table 2. Average top-1 and top-10 accuracy, and NDCG-10 for venue prediction.

City	Method	VenuePop	ContentBased	NearestNeigh	FPMC-LR	MFTP
London	Top-1(%)	8.25	11.01	28.71	35.21	38.25
	Top-10(%)	25.31	39.09	49.48	63.81	65.98
	NDCG-10(%)	15.95	23.01	38.29	48.68	51.16
Singapore	Top-1(%)	13.38	12.76	19.13	31.27	32.50
	Top-10(%)	39.35	41.29	41.59	69.60	71.66
	NDCG-10(%)	25.04	24.86	29.14	48.84	50.52
NewYork	Top-1(%)	11.03	10.07	27.72	38.99	40.96
	Top-10(%)	25.79	37.72	46.60	65.25	66.97
	NDCG-10(%)	17.26	21.86	36.52	51.08	52.91

- **ContentBased.** This method is solely based on the content of social images. Due to the fact that the number of images for unpopular venues is not enough for training the SVM classifiers, we simply calculate the cosine similarity between the venue features and the given social image feature.
- **NearestNeigh.** According to our statistics, users usually move within a bounded region. Therefore, for each social image, we select the nearest neighbor venue to the user's last check-in as the venue tag.
- **FPMC-LR** [8]**.** This method mainly exploits the personalized Markov chains in the inter check-in sequence. The difference from our model lies in that FPMC-LR predict the specific venue in one step and overlooks the importance of the periodical patterns.

Parameter Settings. We randomly initialized model parameters with a Gaussian distribution (with a mean of 0 and standard deviation of 0.01) and optimized the model with stochastic gradient descent (SGD) until convergence. Finally, we tested the batch size of 32, the latent feature dimension of 32, the learning rate of [0.01, 0.05] and the regularizer of 0.01.

Table 2 shows the performance for different models. The results show that:

- The general trend is that MFTP significantly outperforms the rule-based baselines, such as VenuePop and NearestNeigh. Despite the relatively bad performance of NearestNeigh method, it is still better than expected, especially on TOP-1 accuracy. This maybe due to the fact that after data cleaning there are only about 10,000 venues left in each city, which are more likely to be scattered throughout the city and thus have relatively large inter distances. Moreover, as users tend to move within a bounded region, the large inter distance would narrow down the number of venue candidates and hence boost the performance of venue prediction. Therefore, we hypothesize that NearestNeigh would work well on datasets with low density of venues.
- The performance of content-based method ContentBased is unsatisfactory, due to the high diversity of content of social images. For the same venue, people may take photos of different objects or from different angles. For example,

it is common that users would take photo of a dish, upload it to the social network and then check-in to the restaurant. In such a case, the visual content cannot reflect the unique characteristics of the specific restaurant.
– MFTP shows superiority over the strong baseline—FPMC-LR, which is based on a similar generic embedding framework. The possible reasons are as follows: (1) FPMC-LR directly predicts the specific venue tags in which the category of the venue is not utilized; (2) although FPMC-LR considers the transition patterns between successive check-ins, it ignores the important periodical patterns of users' check-in behaviors; and (3) FPMC-LR is a POI recommendation framework in which no visual content information is utilized.

4.3 Illustrative Examples

To gain insights on the influential factors on the task of venue prediction, we comparatively illustrate a few representative examples in Fig. 3. From this figure, we have the following observations:

– The image in Fig. 3(a) refers to a famous landmark in Singapore and the visual content is clean and distinctive. In such cases, the visual content will dominate the prediction.
– From the visual content of the image in Fig. 3(b), we can easily tell that the image was taken in some *library*. However, it is difficult to figure out which specific *library* it refers to. Fortunately, the user's last check-in was to a *bar* which is quite near the specific *library*; this helps us to get the right answer.

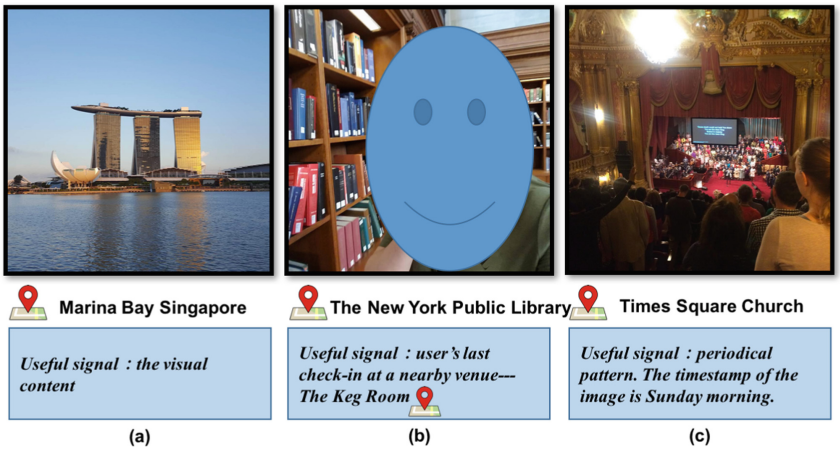

Fig. 3. Illustration of prediction results. They respectively justify the importance of visual content, transition patterns on successive check-ins, and periodical patterns.

– The image in Fig. 3(c) reflects a *church* in New York. However, the visual signals derived from CNNs mistakenly indicate that this image is a *concert hall*. Fortunately, we found that the time stamp of this image is *Sunday moring* and the user periodically visit *church* around the same time every week. Based on these temporal signals, the model could correctly generate the prediction.

Due to space limitation, we only show the positive results in Fig. 3. It is actually a very hard problem to get the specific venue prediction right in some cases. For example, for users that travel frequently, the prediction performance would be quite low. To tackle such problems, we will investigate how to include GPS information to extend our model for further improvement.

5 Conclusion

In this work, we studied the novel problem of specific venue prediction of social images. We first conducted exploratory analysis on real-world datasets, based on which we found strong evidence of transition patterns on successive check-ins and periodical patterns on venue categories. We then developed a generic embedding model based on matrix factorization to capture the interactions between visual content and temporal patterns. To the best of our knowledge, this is the first work on time-aware social image venue prediction. Experimental results on a real-world dataset demonstrate the effectiveness of our proposed solution, where the accuracy of venue prediction was improved by more than 5% by leveraging LBSN check-ins. Apart from quantitative analysis, we highlight two qualitative insights gained from this work. First, it is promising to exploit the venue category information for location-related tasks. Second, transition patterns and periodical patterns are strong signals in predicting users' movements and activities. In future, we plan to investigate the effect of GPS information for venue prediction of multimedia content. Further, we are interested in exploring the recently developed neural factorization machines [14] for modelling the higher-order interactions between users, venues, and temporal patterns.

References

1. Avrithis, Y.S., Kalantidis, Y., Tolias, G., Spyrou, E.: Retrieving landmark and non-landmark images from community photo collections. In: MM. ACM (2010)
2. Cao, S., Snavely, N.: Graph-based discriminative learning for location recognition. IJCV **112**(2), 239–254 (2015)
3. Chen, D.M., Baatz, G., Köser, K., Tsai, S.S., Vedantham, R., Pylvänäinen, T., Roimela, K., Chen, X., Bach, J., Pollefeys, M., Girod, B., Grzeszczuk, R.: City-scale landmark identification on mobile devices. In: CVPR. IEEE (2011)
4. Chen, J., Ngo, C.: Deep-based ingredient recognition for cooking recipe retrieval. In: MM. ACM (2016)
5. Chen, J., Song, X., Nie, L., Wang, X., Zhang, H., Chua, T.: Micro tells macro: predicting the popularity of micro-videos via a transductive model. In: MM, pp. 898–907. ACM (2016)

6. Chen, J., Zhang, H., He, X., Nie, L., Liu, W., Chua, T.: Attentive collaborative filtering: multimedia recommendation with item- and component-level attention. In: SIGIR, pp. 335–344. ACM (2017)
7. Chen, T., He, X., Kan, M.: Context-aware image Tweet modelling and recommendation. In: MM. ACM (2016)
8. Cheng, C., Yang, H., Lyu, M.R., King, I.: Where you like to go next: successive point-of-interest recommendation. In: IJCAI. IJCAI/AAAI (2013)
9. Crandall, D.J., Backstrom, L., Huttenlocher, D.P., Kleinberg, J.M.: Mapping the world's photos. In: WWW. ACM (2009)
10. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: a large-scale hierarchical image database. In: CVPR. IEEE (2009)
11. Farseev, A., Nie, L., Akbari, M., Chua, T.: Harvesting multiple sources for user profile learning: a big datastudy. In: ICMR. ACM (2015)
12. Hays, J., Efros, A.A.: IM2GPS: estimating geographic information from a single image. In: CVPR. IEEE Computer Society (2008)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. IEEE Computer Society (2016)
14. He, X., Chua, T.-S.: Neural factorization machines for sparse predictive analytics. In: SIGIR, pp. 355–364 (2017)
15. He, X., Gao, M., Kan, M.-Y., Liu, Y., Sugiyama, K.: Predicting the popularity of web 2.0 items based on user comments. In: SIGIR, pp. 233–242 (2014)
16. He, X., Liao, L., Zhang, H., Nie, L., Hu, X., Chua, T.-S.: Neural collaborative filtering. In: WWW, pp. 173–182 (2017)
17. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. TOIS **20**(4), 422–446 (2002)
18. Li, X., Pham, T.N., Cong, G., Yuan, Q., Li, X., Krishnaswamy, S.: Where you Instagram? Associating your Instagram photos with points of interest. In: CIKM. ACM (2015)
19. Li, Y., Crandall, D.J., Huttenlocher, D.P.: Landmark classification in large-scale image collections. In: ICCV. IEEE (2009)
20. Zhang, H., Shen, F., Liu, W., He, X., Luan, H., Chua, T.-S.: Discrete collaborative filtering. In: SIGIR, pp. 325–334 (2016)
21. Zhang, J., Nie, L., Wang, X., He, X., Huang, X., Chua, T.: Shorter-is-better: venue category estimation from micro-video. In: MM. ACM (2016)
22. Zhou, B., Lapedriza, À., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: NIPS (2014)