Photo2Trip: Exploiting Visual Contents in Geo-Tagged Photos for Personalized Tour Recommendation

Pengpeng Zhao, Chengfeng Xu, Yanchi Liu, Victor S. Sheng, *Senior Member, IEEE*, Kai Zheng, *Member, IEEE*, Hui Xiong, *Senior Member, IEEE*, and Xiaofang Zhou, *Fellow, IEEE*

Abstract—Recently accumulated massive amounts of geo-tagged photos provide an excellent opportunity to understand human behaviors and can be used for personalized tour recommendation. However, no existing work has considered the visual content information in these photos for tour recommendation. We believe the visual features of photos provide valuable information on measuring user / Point-of-Interest (POI) similarities, which is challenging due to data sparsity. To this end, in this paper, we propose a visual feature enhanced tour recommender system, named 'Photo2Trip', to utilize the visual contents and collaborative filtering models for recommendation. Specifically, we propose a Visual-enhanced Probabilistic Matrix Factorization model (VPMF), which integrates visual features into the collaborative filtering model, to learn user interests by leveraging the historical travel records. We then extend VPMF to End-to-End training framework to incorporate users (POIs) latent factors into the learning process of the visual content of photos, which generalizes the applicability of the proposed VPMF framework in tour recommendation. Extensive empirical studies verify that our proposed visual-enhanced personalized tour recommendation method outperforms other benchmark methods in terms of recommendation accuracy. The results also show that visual features are effective in alleviating the data sparsity and cold start problems on personalized tour recommendation.

Index Terms—Tour recommendation, collaborative filtering, visual content

1 Introduction

Recent years have witnessed a revolution in location-based social network (LBSN) services. As a consequence, large amounts of geo-tagged photos have been accumulated from users. These footprints (or check-ins) provide an excellent opportunity to understand human behaviors and can be used in many fields, including personalized tour recommendation. Tour recommendation aims to find a trip route visiting several POIs that maximize the utility of users according to their trip constraints and their specific interests on POIs. Moreover, it can help tourists narrow down candidate POIs to visit, and plan an appropriate visit order and corresponding duration at each POI in an unfamiliar place.

Tour recommendation and itinerary planning are challenging tasks because tourists have different interests and trip constraints, such as time limitation, the popularity of POIs, and travel time between POIs [15]. Therefore, how to learn user specific interests plays an important role in personalized tour recommendation. Brilhante et al. [5], [6] used visit frequency in a POI category as user visit preference. Lim et al. [25] used average visit duration of all users in a POI category as user interest and took personal visit duration into consideration in tour recommendation, which got better results than frequency-based approaches. However, if a user has not visited any POIs in a category yet, the above methods are not able to make personalized tour recommendation. A straightforward solution is leveraging collaborative filtering to predict user interest of each unvisited POI.

Nevertheless, the check-in data of LBSN is extremely sparse since most users are not residents in their tour destinations. And the sparsity issue causes difficulties for collaborative filtering methods to learn effectively. Besides, the cold start problem (no historical check-in records for new users or new POIs) is even more severe in personalized tour recommendation. Therefore, additional information needs to be incorporated to address these issues. We find that the visual features in geo-tagged photos taken by users can provide important context information for predicting user visit interests. From these photos, the POI information can be inferred, also users' behaviors and preferences can be

- *P. Zhao and C. Xu are with the Institute of AI, School of Computer Science and Technology, Soochow University, Suzhou 215006, China. E-mail: ppzhao@suda.edu.cn, cfxu@stu.suda.edu.cn.*
- *Y. Liu is with the NEC Labs America, Princeton, NJ 08540 USA. E-mail: yanchi@nec-labs.com.*
- *V.S. Sheng is with the Department of Computer Science, Texas Tech University, Lubbock, TX 79409 USA. E-mail: victor.sheng@ttu.edu.*
- *K. Zheng is with the University of Electronic Science and Technology, Chengdu 610054, China. E-mail: zhengkai@uestc.edu.cn.*
- *H. Xiong is with the Management Science and Information Systems Department, Rutgers University, Piscataway, NJ 08854 USA. E-mail: hxiong@rutgers.edu.*
- *X. Zhou is with the University of Queensland, Brisbane, QLD 4072, Australia. E-mail: zxf@itee.uq.edu.au.*

Manuscript received 25 Dec. 2018; revised 6 July 2019; accepted 15 Sept. 2019. Date of publication 26 Sept. 2019; date of current version 5 Mar. 2021.
(Corresponding author: Pengpeng Zhao.)
Recommended for acceptance by M. Wang.
Digital Object Identifier no. 10.1109/TKDE.2019.2943854

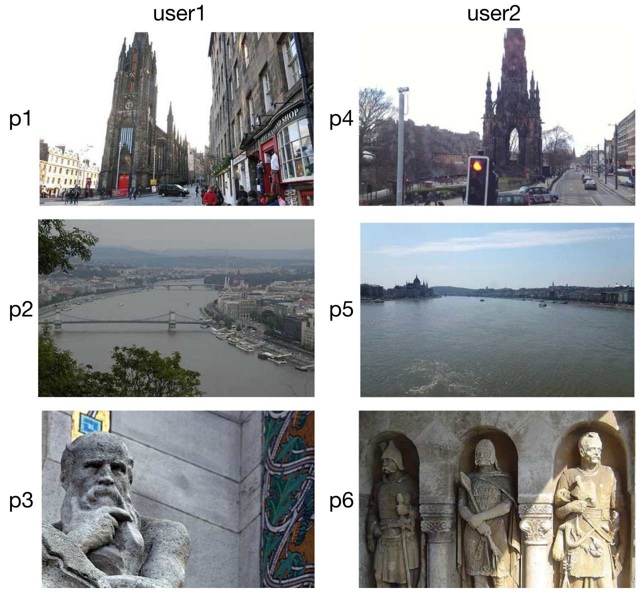


Fig. 1. Three pairs of POI photos from six different POIs and visited by two users having similar visual appearances.

revealed. For example, Fig. 1 shows three pairs of POI photos from two different users. A tourist who favors the POIs in the first column might also be interested in the second one since they exhibit similar visual appearances. These observations motivate us to leverage the visual information, which is overlooked by existing methods, in addition to others for personalized tour recommendation.

In this paper, we propose a solution that leverages the visual contents of geo-tagged photos together with collaborative filtering models for personalized tour recommendation. Specifically, we first extract various visual features from photos taken by tourists, and utilize them to understand the styles of the POIs and the visual preferences of users. Then, we propose a Visual-enhanced Probabilistic Matrix Factorization model (VPMF), which integrates visual features into the collaborative filtering model, to learn user interests by leveraging the historical travel records of peer users. After that, user interests together with trip constraints are formalized to an optimization problem for trip planning. Compared with the state-of-the-art methods, our proposed model consistently improves the performance of visit interest prediction for tour recommendation.

Although our previous work [49] has already solved the cold start problem by utilizing the visual features of photos, it fails to involve users' (POIs') latent factors in the learning process, while they are critical to improve the performance of tour recommendation. Images uploaded by users and associated with POIs contain rich information about user preferences and POI properties. A user's latent features should be differentiated based on whether an image is posted by the user or not. The same is true for the latent vectors of POIs. To address this issue, in Section 4, we propose an End-to-End VPMF (E2E-VPMF) training framework to incorporate users' (POIs') latent factors of the collaborative filtering model into the learning process of the visual content of photos. In this way, E2E-VPMF generalizes the applicability of the proposed VPMF framework in tour recommendation. Finally, we can obtain more specific users' preferences and POIs' representations to achieve more accurate POI tour recommendations. Our experimental results on real-world Yahoo! Flickr dataset show that E2E-VPMF significantly outperforms VPMF and all other baselines. On average, it improves over 7.21 percent on trip planning with respect to F_1 and over 32.92 percent on visit duration prediction with respect to Root Mean Square Error ($RMSE$), comparing with the strongest baseline PersTour.

To summarize, our new technical contributions in this extension are listed as follows.

- To the best of our knowledge, this is the first work that utilizes visual features of geo-tagged photos to learn user interests for personalized tour recommendation.
- A VPMF model is proposed to integrate visual features into the collaborative filtering model to enhance its performance. The model uses the content of user-generated photos to improve the prediction accuracy. Moreover, it reduces the negative impacts of the data sparsity problem and the cold start problem.
- In addition, an End-to-End VPMF (E2E-VPMF) method is further proposed to incorporate an end-to-end training framework with visual contents into the VPMF method. E2E-VPMF is able to use the visual features to guide the learning process of users' (POIs') latent factors.
- Extensive experiments are conducted to study the impact of the key parameters and the effectiveness of our newly proposed method in terms of different metrics, such as precision, recall, F_1, and RMSE.

The rest of this paper is organized as follows. Section 2 introduces the problem definition and preliminaries. The system framework and the proposed visual feature recommendation algorithm E2E-VPMF are presented in Section 3. We report the experimental results and discussions in Section 4. Section 5 surveys the related work and Section 6 concludes this paper.

2 Preliminaries

In this section, we first introduce some basic concepts of tour recommendation and then give the formal problem definition, followed by the correlation analysis between the visual features of POIs and users' ratings. At last, we introduce a basic collaborative filtering model and three visual features will be used in our personalized tour recommendation system.

2.1 Basic Concepts

Popularity. The popularity of a POI is defined as the number of times that the POI has been visited.

Time-Based User Interest. We define the interest of a user in a POI as the ratio between the personal visit duration and the average visit duration of all users.

Personalized POI Visit Duration. With the definition of time-based user interest, we can define the personalized visit duration of POI as the multiplication of user interest and the average time spent at POI.

Travel Time. Travel time is the time cost moving from one POI to another, which is based on the distance between two POIs and and the given moving speed.

TABLE 1
List of Notations

Notation	Explanation
$\mathcal{U},\ \mathcal{L},\ \mathcal{P}$	sets of users, POIs and photos
N, M, S	the number of users, POIs and photos
u_*, l_*, p_*	a user, a POI and a photo
$\mathbf{X}$	the check-in frequency matrix
$\mathbf{R}$	the normalized version of $\mathbf{X}$
$\mathbf{U}$, $\mathbf{V}$	the latent feature matrices of users and POIs
D	dimensionality of the latent vector
K	dimensionality of the visual feature vector
s(*,*)	the similarity of two POIs or two users
N_*	neighborhoods of a POI or a user
$Cat(*), Pop(*)$	the category and the popularity of a POI
$Int(*)$	time-based user interest in a POI
$Cost(*,*)$	travel cost between two POIs
$Vis(*)$	the visual feature of a photo
$Pro(*)$	the probability that a photo is posted by an user
$\mathcal{H}$	set of photos posted by an user
$\mathcal{W}$	set of photos associated with a POI
$\mathcal{J},\ \mathcal{L}$	the objective function
P, Q	the interaction matrix
L	the number of negative sample for each photo
$\sigma_*^2,\ \lambda_*$	the variance and the regularization term
α	the balancing parameter

Travel Cost. The cost of traveling from one POI to another is calculated as the summation of the travelling time and the personalized visit duration of POI.

2.2 Problem Formulation

In this subsection, we introduce notations used in this paper. Let $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ be a set of N users, $\mathcal{L} = \{l_1, l_2, \ldots, l_M\}$ be a set of M POIs (locations), and $\mathcal{P} = \{p_1, p_2, \ldots, p_S\}$ be a set of S photos. $\mathbf{X} \in \mathbb{R}^{N\times M}$ denotes a user-POI check-in matrix, where $\mathbf{X}_{ij}$ represents the frequency that u_i checked in l_j. We denote the normalized version of $\mathbf{X}$ as $\mathbf{R} \in \mathbb{R}^{N\times M}$, where $\mathbf{R}_{ij} = g(\mathbf{X}_{ij})$ and $g(x) = \frac{1}{1+exp^{-x}}$. In addition, users can post photos to LBSNs and add locations to the photos. $\mathcal{P}_{u_i}$ represents a set of photos posted by u_i. $\mathcal{P}_{l_j}$ represents a set of images that are associated with l_j.

Then our task is formally stated as: Given a user u with a starting POI l_1, an ending POI l_n and a time budget B, we need to find an optimal trip route $I = (l_1, \ldots, l_n)$ that maximizes user utility under following constraints: (1) it starts at location l_1 and ends at location l_n; (2) it completes within the time budget B. The utility of visiting a POI l_i is represented by the popularity and the user interest of this POI, which are denoted as $Pop(l_i)$ and $Int(l_i)$, respectively. The traveling from l_i to l_j is calculated as the summation of the traveling time and the personalized visit duration of the POI l_j. The involved notations and their definitions are listed in Table 1 for clarity.

2.3 Correlation Analysis

Before designing a tour recommendation model, it is important to understand tourist visit behaviors. In other words, we try to answer the question: "do tourist visit behaviors correlate with the visual style and appearance of POIs?" To answer this question, we analyzed the correlation between visual contents in photos of POIs and time-based user interest. Our analysis results are shown in Fig. 2, in which $RMSE$ metric is used to measure the prediction error and the smaller value is the better. First, we predict user personalized POI visit duration using the average visit duration of all users of the category of the POI. Then given $top-k$ most similar POIs on visual appearance as neighborhoods of a POI, we predict user visit duration at the POI using the average visit duration of its neighborhoods. Finally, given $top-k$ most similar users on the visual content of photos posted by users as neighborhoods of a user, we predict user visit duration at the POI using the average visit duration of the POI taken by his/her neighborhoods. The above three operations correspond to the three legends in Fig. 2, respectively. From Fig. 2, we can see that compared to use the category of a POI (i.e., the first legend), the prediction error is reduced either under given visually similar neighbors of a POI (i.e., the second legend) or under given neighbors of a user with similar visual taste (i.e., the third legend). In both cases, the neighbors of POIs and users are selected from images with the similar visual content. And the value of RMSE is reduced about 6.3 and 2.4 percent on average, which indicates that the visual contents in photos of POIs are effective for capturing time-based user interest. Therefore, we can see the answer to the above question is "yes".

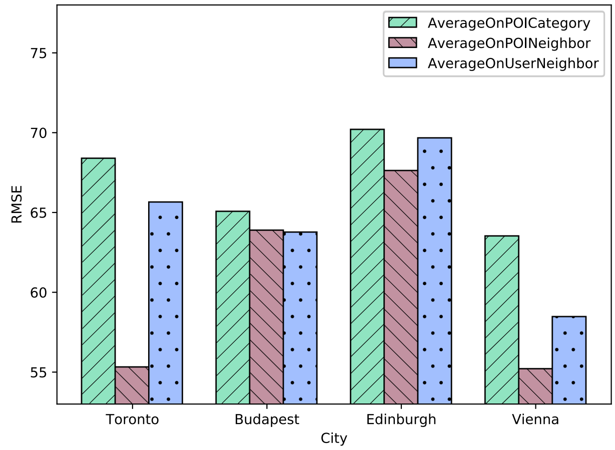


Fig. 2. The effects of user interest prediction under given visual neighborhoods of POIs and users, respectively.

2.4 PMF Model

Probabilistic Matrix Factorization (PMF) [35] is a simple, accurate, and efficient model among collaborative filtering methods and has been widely adopted for POI recommendation [10], [28]. PMF not only can deal with very large datasets, but also has the ability to make recommendations for users with only a few ratings in recommender systems [16], [42]. We will later show how to improve PMF with visual features in Section 4. The likelihood of observing a specific user-POI relation in $\mathbf{R}$ can be expressed as follows.

$$p(\mathbf{R}|\mathbf{U}, \mathbf{V}, \sigma^2) = \prod_{i=1}^{N}\prod_{j=1}^{M}\left[\mathcal{N}(\mathbf{R}_{ij}|\mathbf{U}_i^T\mathbf{V}_j, \sigma^2)\right]^{\mathbf{Y}_{ij}} \tag{1}$$

$$p(\mathbf{U}) = \mathcal{N}(0, \sigma_U^2),\ \ p(\mathbf{V}) = \mathcal{N}(0, \sigma_V^2), \tag{2}$$

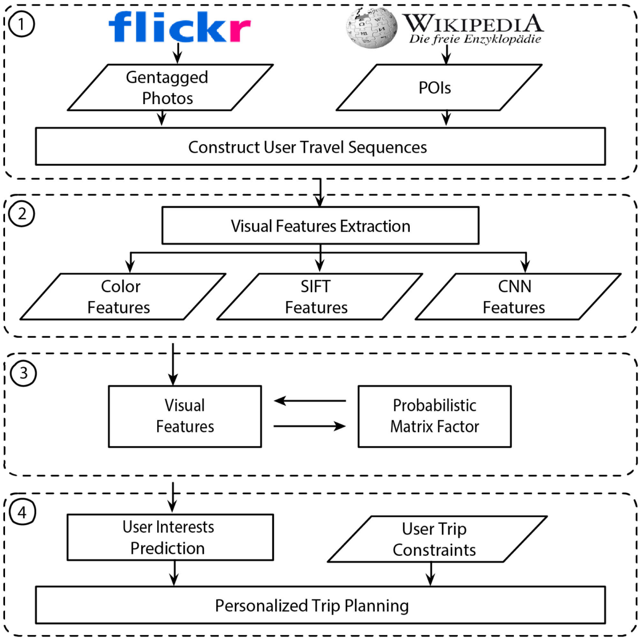


Fig. 3. Framework of Photo2Trip recommender system.

where $\mathcal{N}(x|\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ, and $\mathbf{Y}$ is an indicator matrix, in which $\mathbf{Y}_{ij} = 1$ if $\mathbf{R}_{ij} > 0$ and 0 otherwise. The observation R is modeled as a draw from a Gaussian distribution, where the mean of R_{ij} is $U_i^T V_j$ and the variance is σ. $\mathbf{U} \in \mathbb{R}^{D\times N}$ and $\mathbf{V} \in \mathbb{R}^{D\times M}$ are the latent feature matrices of users and POIs, which are also drawn from the zero-mean normal distribution.

Now, through a Bayesian inference, we can obtain the posterior probability of $\mathbf{U}$ and $\mathbf{V}$ as follows.

$$p(\mathbf{U},\mathbf{V}|\mathbf{R},\sigma^2,\sigma^2_{\mathbf{U}},\sigma^2_{\mathbf{V}}) \propto p(\mathbf{R}|\mathbf{U},\mathbf{V},\sigma^2)p(\mathbf{U}|\sigma^2_U)p(\mathbf{V}|\sigma^2_V). \quad (3)$$

To calculate $\mathbf{U}$ and $\mathbf{V}$, so as to maximize the posterior probability given observation $\mathbf{R}$, we can learn the latent feature $\mathbf{U}$ and $\mathbf{V}$ of users and POIs purely based on the observation $\mathbf{R}$ using Equation (3).

2.5 Visual Features in Geo-Tagged Photos

There are lots of different types of visual features in geo-tagged photos. In order to improve recommendation accuracy, we should choose visual features in a proper way. We assume that tourists are attracted by the visual effects of POIs, such as colors, abstract features, and visual contents, as shown in Fig. 1. More specifically, given two POIs l_i and l_j, we could calculate the similarity $s(l_i, l_j)$ between the two POIs by measuring their visual correlation through extracted visual features. Next we introduce some widely used visual features.

Color Histogram. In POI photos, color is the first impression to people. For example, POI photos with large color areas, such as blue sky, golden beaches, and blue sea water, have a deep impression on users. Color histogram is a widely used visual feature. We adopt a standard color histogram feature and extract a 512-dimensional color feature vector for each photo. And a joint histogram in RGB color space has 8 bins in each channel.

Scale-Invariant Feature Transform (SIFT). For point description, the SIFT descriptor [31] is known as scale-invariant features and widely used in object recognized and content-based image search for its good classification accuracy [11]. The SIFT finds interest points and captures the local shape around it using edge orientation histograms. SIFT features are also robust to changes in lighting, noise, and minor differences in viewpoint. Because many of the photos are taken from the same scene but different angles, SIFT will be useful in this scenario. We extract a 128-dimensional SIFT feature after resizing each POI photo to 256 × 256 pixels.

Convolutional Neural Networks. Different from above hand-crafted visual features, convolutional neural network (CNN) can automatically discover high-level visual features of photos by learning from training data. It has been shown that CNN performs well in image classification and object detection. The features extracted by CNN can reflect a photo globally, regionally, and locally. Intuitively, these features (or some of them) should be useful for visual recommendation as we will show in our later experiments. In this paper we use the VGG16 model [36], which is the state-of-the-art architecture, to extract features from user-generated geo-tagged photos. It is composed of 13 convolution, 5 max pooling, 3 fully connected and 1 softmax layers. Specifically, we resize each photo to 224 × 224 pixels as the input of VGG16 and obtain a 4096 dimension visual feature vector as the output of the second fully-connected layer.

3 Photo2Trip Tour Recommendation

3.1 System Framework Overview

Fig. 3 shows the overall framework of our Photo2Trip personalized tour recommendation system, which is composed of four main parts. First, we crawl photos from a public photo-sharing web site (i.e., Flickr). With the same approach described in [23], we obtain a list of POIs from Wikipedia and map these photos to user-POI visits. And we construct user travel sequences based on them. Second, after mining the travel patterns of users' trip sequences, we extract the visual features in the user-generated photos using the visual toolbox. Next, we propose an end-to-end visual probability matrix factorization (E32E-VPMF) model to predict user visit interests, where the visual features of images influence the learning of probabilistic matrix factorization model. In other words, the extraction of visual contents of photos guides the learning process of users and POIs latent features. Last, with user's input trip constraints, including travel time limitation, the starting POI, and the ending POI, the trip planning module generates a personalized trip route that maximizes user utility while adhering to the user's trip constraints. Trip planning is further modeled as an orienteering problem and solved using linear programming.

In the following subsections, we will introduce three essential modules (i.e., user interest prediction and personalized trip planning) in our framework in detail. Before that, we first briefly describe the basic visual-enhanced probabilistic matrix model.

3.2 VPMF Model

As observed in Section 2.3, user visit behaviors are related to the visual appearance of POIs, and the visual contents in user-posted photos reflect the user visit preferences. According to the idea of neighbor-based collaborative filtering, it is natural to assume that the visit behavior and the visual taste of a user are similar to that of his/her neighbors, and the interests of a POI are similar to those of its similar visual POIs. Based on the above analysis, we propose a visual-enhanced PMF model to improve user interest prediction accuracy. We first select $top-k$ nearest neighbors for each POI and for each user respectively based on the visual content similarity of the photos of POIs and the photos taken by users, where k is empirically chosen according different datasets. And then we incorporate the constructed visual neighborhoods into the learning process of PMF.

Since each POI has more than one photo, to get a representative visual features vector of a POI, we merge each dimension visual vector extracted from POI photos using a maximum pooling method. After that, we linearly combine three similarities of different visual features to represent the final feature vector of the POI. Then the similarity of two POIs is measured by the cosine similarity of the visual feature vectors, which is denoted as $s(l_i, l_j)$. The similarity $s(u_i, u_j)$ of two users is also calculated by the cosine similarity of the visual vectors of photos posted by the users in the same way. There are more approaches fusing multiple visual features, such as multi-modal graph-based reranking [38] and non-linear feature fusion [39].

Inspired by neighborhood MF [21], in the probability matrix factorization process, the latent features of users u_i and POIs l_j should be close to their neighborhoods N_{u_i} and N_{l_j} respectively. Based on this intuition, we add Gaussian priors to user's and POI's latent feature vectors to ensure that $\mathbf{U}_i$ and $\mathbf{V}_j$ are centered around the mean of their neighborhood and formulate the following equations.

$$\mathbf{U}_i = \sum_{t\in N_{u_i}} s(i,t) \times \mathbf{U}_t + \tilde{\mathbf{U}}_i, \quad \tilde{\mathbf{U}}_i \sim \mathcal{N}(0, \sigma_U^2 \mathbf{I}) \tag{4}$$

$$\mathbf{V}_j = \sum_{t\in N_{l_j}} s(j,t) \times \mathbf{V}_t + \tilde{\mathbf{V}}_j, \quad \tilde{\mathbf{V}}_j \sim \mathcal{N}(0, \sigma_V^2 \mathbf{I}). \tag{5}$$

In the above two equations, the latent feature vector of each user and each POI comprise of two terms. The first term characterizes the neighborhood related feature of the user or the POI. For notation convenience, we normalize the similarities to ensure $\sum_{t\in N_{u_i}} s(i,t) = 1$ and $\sum_{t\in N_{plj}} s(j,t) = 1$. The second term emphasizes the unique feature of each user and each POI, which could diverge from their neighborhood. The variance parameter σ_U^2 and σ_V^2 are used to control the divergence. The lower the variance, the less diverges the feature vector from that of the neighbors. With the visual neighborhood incorporated, the conditional distributions of the observed $\mathbf{R}$, as shown in Equation 1, does not change. Based on the Bayesian formula, the posterior distribution over the latent factors of users and POIs is given as follows (Equation (6)).

$$\begin{aligned} p(\mathbf{U}, \mathbf{V} | \mathbf{R}, \sigma^2, \sigma_U^2, \sigma_V^2) = & \prod_{i=1}^{N} \prod_{j=1}^{M} [\mathcal{N}(\mathbf{R}_{ij} | \mathbf{U}_i^T \mathbf{V}_j, \sigma^2)]^{\mathbf{Y}_{ij}} \\ & \times \prod_{i=1}^{N} \mathcal{N}(\mathbf{U}_i | \sum_{t\in N_{u_i}} s(i,t) \times \mathbf{U}_t, \sigma_U^2 \mathbf{I}) \\ & \times \prod_{j=1}^{M} \mathcal{N}(\mathbf{V}_j | \sum_{t\in N_{l_j}} s(j,t) \times \mathbf{V}_t, \sigma_V^2 \mathbf{I}). \end{aligned} \tag{6}$$

3.3 The Proposed Framework

For the VPMF model proposed above, the visual features of the pictures are unchanged in the training process. That is, the extracted visual features are loaded as a hyper-parameter into the model. Enlightened by [41], we propose an extended version of VPMF, namely End-to-End VPMF(E2E-VPMF) to further improve the performance of tour recommendation. In this subsection, we first model the visual content of images, and then introduce the framework we proposed and the process of negative sampling in detail. At last, we use gradient descent to update the variables alternatively.

3.3.1 Extracting Visual Features

To utilize the visual features of photos in tour recommendation, we first need to extract valuable visual contents from photos taken by users or POIs. Convolutional Neural Network is a powerful deep neural network of discovering high-order visual characteristics of photos for various applications. Thus, we apply VGG16 framework of CNN to extract features from photos as described in Section 2.5. We denote the visual content extracted as a feature learning function $Vis(p_k)$, since the weights of CNN will be constantly updated to direct and involve in the learning process of latent factors of each user/POI. As usual, we will not train VGG16 model from the ground up. Instead, we use the pre-trained VGG16 framework for time and space complexity [34]. Through the photos contents extracted by VGG16, we will combine these contents with the proposed model for tour recommendation.

Given a photo p_r posted by u_i, it is natural to assume that u_i's interests are associated with certain visual contents in p_r; yet for an arbitrary photo p_f posted by other users, i.e., $p_f \notin \mathcal{P}_{u_i}$, u_i's interests are less likely to be associated with certain visual contents in p_f. Meanwhile u_i's interests are now characterized by the latent factor $\mathbf{U}_i$. It signifies that $\mathbf{U}_i$ should be able to distinguish whether a photo p_r is posted by u_i based on the visual feature $Vis(p_r)$. Thus, we denote the probability that p_r is posted by u_i as $Pro(h_{ir} = 1 | u_i, p_r)$, where h_{ir} denotes if p_r is posted by u_i or not. $Pro(h_{ir} = 1 | u_i, p_r)$ is given as

$$Pro(h_{ir} = 1 | u_i, p_r) = \frac{exp(\mathbf{U}_i^T \cdot \mathbf{P} \cdot Vis(p_r))}{\sum_{p_k \in \mathcal{P}} exp(\mathbf{U}_i^T \cdot \mathbf{P} \cdot Vis(p_k))}, \tag{7}$$

where $\mathbf{P} \in \mathbb{R}^{D\times K}$ is an interaction matrix between the visual feature and the user' latent factor, and K is the dimension size of visual feature vector, where K=4096 by the output of VGG16 with the last two layers removed. Hence, for $p_r \in \mathcal{P}_{u_i}$, by maximizing $Pro(h_{ir} = 1 | u_i, p_r)$, we can force $\mathbf{U}_i$ to move toward the visual imagery by the matrix $\mathbf{P}$. In such a manner, the visual contents of photos can direct the optimizing process of $\mathbf{U}_i$.

Similarly, given a photo p_t attached with l_j, the visual contents of p_t are likely to describe location l_j ; while for an arbitrary photo p_f that is not attached with l_j, the visual content of p_f is less likely to describe l_j. At the same time, l_j is now described by the latent factor $\mathbf{V}_j$, where $\mathbf{V}_j$ should be able to distinguish whether a photo p_t is attached with l_j based on the visual feature $Vis(p_t)$. In the same way, we denote the probability that p_t is attached with l_j as $Pro(w_{jt} = 1 | l_j, p_t)$,

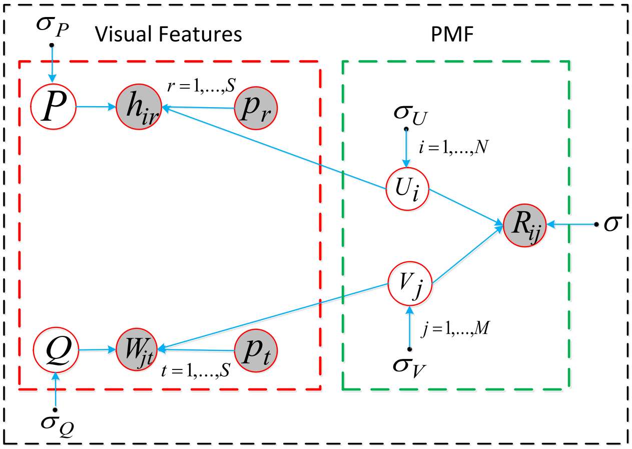


Fig. 4. The graphical representation of the E2E-VPMF model.

where w_{jt} denotes if p_t is attached with l_j or not. $Pro(w_{jt}=1|l_j,p_t)$ is written as

$$Pro(w_{jt}=1|l_j,p_t)=\frac{exp(\mathbf{V}_j^T\cdot\mathbf{Q}\cdot Vis(p_t))}{\sum_{p_k\in\mathcal{P}}exp(\mathbf{V}_j^T\cdot\mathbf{Q}\cdot Vis(p_k))}, \tag{8}$$

where $\mathbf{Q}\in\mathbb{R}^{D\times K}$ is an interaction matrix between the visual feature and the POI's latent factor. Therefore, for $p_t\in\mathcal{P}_{l_j}$, by maximizing $Pro(w_{jt}=1|l_j,p_t)$, we compel $\mathbf{V}_j$ to approach the visual imagery by the matrix $\mathbf{Q}$. In such a manner, the visual contents of photos can also guide the optimizing process of $\mathbf{V}_j$.

In the above two equations, $Pro(h_{ir}=1|u_i,p_r)$ and $Pro(w_{jt}=1|l_j,p_t)$ are defined as the probability that the user and the POI are associated with photos, respectively. The large probability indicates that the more the photos fit the user's preference and the more relevant the photos are to a POI. Thus, the likelihood function of modeling visual features of photos is given as follows:

$$\begin{aligned}Pro(\mathcal{H},\mathcal{W}|\mathbf{U},\mathbf{V},\mathbf{P},\mathbf{Q})=&\left[\prod_{i=1}^{N}\prod_{p_r\in\mathcal{P}_{u_i}}Pro(h_{ir}=1|u_i,p_r)\right]\\&\times\left[\prod_{j=1}^{M}\prod_{p_t\in\mathcal{P}_{l_j}}Pro(w_{jt}=1|l_j,p_t)\right],\end{aligned} \tag{9}$$

where $\mathcal{H}=\{h_{ir}:u_i\in\mathcal{U}\;\forall\;p_r\in\mathcal{P}_{u_i}\}$ and $\mathcal{W}=\{w_{jt}:l_j\in\mathcal{L}\;\forall\;p_t\in\mathcal{P}_{l_j}\}$. Similarly, $\mathbf{P}$ and $\mathbf{Q}$ are also drawn from the zero-mean Gaussian prior as $p(\mathbf{P}|\sigma_P^2)=\prod_{i=1}^{D}\prod_{j=1}^{K}\mathcal{N}(\mathbf{P}_{ij}|0,\sigma_P^2)$ and $p(\mathbf{Q}|\sigma_Q^2)=\prod_{i=1}^{D}\prod_{j=1}^{K}\mathcal{N}(\mathbf{Q}_{ij}|0,\ \sigma_Q^2)$, where σ_P^2 and σ_Q^2 are the variance.

3.3.2 E2E-VPMF Model

With Equation (3) modeling user-POI relation and Equation (9) modeling the visual features of photos, we propose an end-to-end training framework for tour recommendation based on the Bayesian formula as follows.

$$\begin{aligned}&Pro(\mathbf{U},\mathbf{V},\mathbf{P},\mathbf{Q}|\mathbf{R},\mathcal{H},\mathcal{W},\sigma^2,\sigma_U^2,\sigma_V^2,\sigma_P^2,\sigma_Q^2)\\&\propto Pro(\mathbf{R},\mathcal{H},\mathcal{W}|\mathbf{U},\mathbf{V},\mathbf{P},\mathbf{Q},\sigma^2)\ Pro(\mathbf{U},\mathbf{V},\mathbf{P},\mathbf{Q}|\sigma_U^2,\sigma_V^2,\sigma_P^2,\sigma_Q^2)\\&=Pro(\mathbf{R}|\mathbf{U},\mathbf{V},\sigma^2)\ Pro(\mathcal{H},\mathcal{W}|\mathbf{U},\mathbf{V},\mathbf{P},\mathbf{Q})\\&\quad\times Pro(\mathbf{P})\ Pro(\mathbf{Q})\ Pro(\mathbf{U})\ Pro(\mathbf{V}).\end{aligned} \tag{10}$$

The graphical representation of the proposed framework is shown in Fig. 4. Given these hyper-parameters $\sigma^2,\sigma_U^2,\sigma_V^2,\sigma_P^2,\sigma_Q^2$, maximizing the log posterior is to find $\mathbf{U},\mathbf{V},\mathbf{P},\mathbf{Q},Vis$ in Equation 10 is equivalent to minimizing the following objective function.

$$\begin{aligned}\mathcal{L}=&\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\mathbf{Y}_{ij}}{2}(\mathbf{R}_{ij}-\mathbf{U}_i^T\mathbf{V}_j)^2+\frac{\lambda_1}{2}(||\mathbf{U}||_F^2+||\mathbf{V}||_F^2)\\&-\alpha\sum_{i=1}^{N}\sum_{p_c\in\mathcal{P}_{u_i}}log\ Pro(h_{ic=1}|u_i,p_c)+\frac{\lambda_2}{2}||\mathbf{P}||_F^2\\&-\alpha\sum_{j=1}^{M}\sum_{p_c\in\mathcal{P}_{l_j}}log\ Pro(w_{jc=1}|l_j,p_c)+\frac{\lambda_2}{2}||\mathbf{Q}||_F^2,\end{aligned} \tag{11}$$

where $\lambda_1=\frac{\sigma^2}{\sigma_U^2}=\frac{\sigma^2}{\sigma_V^2}$, $\lambda_2=\frac{\sigma^2}{\sigma_P^2}=\frac{\sigma^2}{\sigma_Q^2}$ to reduce the number of hyper-parameters and α is a balancing parameter.

3.3.3 Negative Sampling

The gradients of $log\ Pro(h_{ic=1}|u_i,p_c)$ and $log\ Pro(w_{jc=1}|l_j,p_c)$ w.r.t $\mathbf{U},\mathbf{V},\mathbf{P},\mathbf{Q}$ involve the calculation of $\sum_{p_k\in\mathcal{P}}exp(\mathbf{U}_i^T\cdot\mathbf{P}\cdot Vis(p_k))$ or $\sum_{p_k\in\mathcal{P}}exp(\mathbf{V}_i^T\cdot\mathbf{P}\cdot Vis(p_k))$, which costs a lot of computational operations and increases time complexity. To simplify the computation, following previous works [33], [40], we take the negative logarithm of likelihood and further approximate $log\ Pro(h_{ic=1}|u_i,p_c)$ as

$$log\ \sigma(\mathbf{U}_i^T\cdot\mathbf{P}\cdot Vis(p_c))+\sum_{s=1}^{L}log\ \sigma(-\mathbf{U}_i^T\cdot\mathbf{P}\cdot Vis(p_{cs})), \tag{12}$$

where we randomly sample L negative samples $p_{cs},s=1,\ldots,L$ for each photo $p_c\in\mathcal{P}_{u_i}$ from photos that are not posted by u_i. Similarly, $log\ Pro(w_{jc=1}|l_j,p_c)$ is approximated as

$$log\ \sigma(\mathbf{V}_j^T\cdot\mathbf{Q}\cdot Vis(p_c))+\sum_{t=1}^{L}log\ \sigma(-\mathbf{V}_j^T\cdot\mathbf{Q}\cdot Vis(p_{ct})), \tag{13}$$

where $p_{ct},t=1,\ldots,L$ are randomly sampled from photos not attached with l_j. In this way, the gradients of variables are simplified by negative sampling , which will be given next.

3.3.4 Update Formulas

To update the variables of the model, we adopt stochastic gradient descent (SGD), which iteratively optimizes a latent variable while fixing the remaining variables. Specifically, the optimal solution of $\mathbf{U}$ (or $\mathbf{V},\mathbf{P},\mathbf{Q}$) can be analytically computed in a closed form by simply differentiating the optimization function $\mathcal{L}$.

The update formula of $\mathcal{L}$ w.r.t $\mathbf{U}_i$ is written as follows.

$$\tfrac{\partial\mathcal{L}}{\partial\mathbf{U}_i}=\sum_{j=1}^{M}\mathbf{Y}_{ij}(\mathbf{R}_{ij}-\mathbf{U}_i^T\mathbf{V}_j)(-\mathbf{V}_j)+\lambda_1\mathbf{U}_i-\alpha A, \tag{14}$$

where $A\in\mathbb{R}^D$ is a vector calculated by substituting Formula (12) into the third term Formula (11) given as

$$\begin{aligned}A=&\sum_{p_c\in\mathcal{P}_{u_i}}[(1-\sigma(\mathbf{U}_i^T\cdot\mathbf{P}\cdot Vis(p_c)))\mathbf{P}\cdot Vis(p_c)\\&-\sum_{s=1}^{L}(1-\sigma(-\mathbf{U}_i^T\cdot\mathbf{P}\cdot Vis(p_{cs})))\mathbf{P}\cdot Vis(p_{cs})].\end{aligned} \tag{15}$$

The update formula of $\mathcal{L}$ w.r.t $\mathbf{V}_j$ is written as

$$\frac{\partial \mathcal{L}}{\partial \mathbf{V}_j} = \sum_{i=1}^{N} Y_{ij}(\mathbf{R}_{ij} - \mathbf{U}_i^T \mathbf{V}_j)(-\mathbf{U}_i) + \lambda_1 \mathbf{V}_j - \alpha B \tag{16}$$

similarly, where $B \in \mathbb{R}^D$ is also a vector calculated by substituting Formula (13) into the fourth term Formula (11) given as

$$\begin{aligned} B = &\sum_{p_c \in \mathcal{P}_{l_j}} [(1 - \sigma(\mathbf{V}_j^T \cdot \mathbf{Q} \cdot Vis(p_c)))\mathbf{Q} \cdot Vis(p_c) \\ &- \sum_{t=1}^{L} (1 - \sigma(-\mathbf{V}_j^T \cdot \mathbf{Q} \cdot Vis(p_{ct})))\mathbf{Q} \cdot Vis(p_{ct})]. \end{aligned} \tag{17}$$

The update formula of $\mathcal{L}$ w.r.t $\mathbf{P}$ can be written as

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{P}} = \alpha &\sum_{i=1}^{N} \sum_{p_c \in \mathcal{P}_{u_i}} [(1 - \sigma(\mathbf{U}_i^T \cdot \mathbf{P} \cdot Vis(p_c)))\mathbf{U}_i \cdot Vis(p_c)^T \\ &- \sum_{s=1}^{L} (1 - \sigma(-\mathbf{U}_i^T \cdot \mathbf{P} \cdot Vis(p_{cs})))\mathbf{U}_i \cdot Vis(p_{cs}^T)] + \lambda_2 \mathbf{P}. \end{aligned} \tag{18}$$

Similarly, the update formula of $\mathcal{L}$ w.r.t $\mathbf{Q}$ can be written as

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{Q}} = \alpha &\sum_{j=1}^{M} \sum_{p_c \in \mathcal{P}_{l_j}} [(1 - \sigma(\mathbf{V}_j^T \cdot \mathbf{Q} \cdot Vis(p_c)))\mathbf{V}_j \cdot Vis(p_c)^T \\ &- \sum_{t=1}^{L} (1 - \sigma(-\mathbf{V}_j^T \cdot \mathbf{Q} \cdot Vis(p_{ct})))\mathbf{V}_j \cdot Vis(p_{ct}^T)] + \lambda_2 \mathbf{Q}. \end{aligned} \tag{19}$$

However, parameter θ of CNN explained below cannot be optimized by an analytic solution as we do for $\mathbf{U}, \mathbf{V}, \mathbf{P}$ and $\mathbf{Q}$ because θ is closely related to the features in CNN architecture. To optimize the parameters for CNN, we use back propagation (BP) algorithm and the partial derivative of $\mathcal{L}$ w.r.t θ, which is given as follows.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \theta} = &\sum_{i=1}^{N} \sum_{p_c \in \mathcal{P}_{u_i}} [(1 - \sigma(\mathbf{U}_i^T \cdot \mathbf{P} \cdot Vis(p_c))) \sum_{h=1}^{K} p_h^T \mathbf{U}_i \frac{\partial Vis(p_c)_h}{\partial \theta} \\ &- \sum_{s=1}^{L} (1 - \sigma(-\mathbf{U}_i^T \cdot \mathbf{P} \cdot Vis(p_{cs}))) \sum_{h=1}^{K} p_h^T \mathbf{U}_i \frac{\partial Vis(p_{cs})_h}{\partial \theta}] \\ &+ \sum_{j=1}^{M} \sum_{p_c \in \mathcal{P}_{l_j}} [(1 - \sigma(\mathbf{V}_j^T \cdot \mathbf{Q} \cdot Vis(p_c))) \sum_{h=1}^{K} q_h^T \mathbf{V}_j \frac{\partial Vis(p_c)_h}{\partial \theta} \\ &- \sum_{s=1}^{L} (1 - \sigma(-\mathbf{V}_j^T \cdot \mathbf{Q} \cdot Vis(p_{ct}))) \sum_{h=1}^{K} q_h^T \mathbf{V}_j \frac{\partial Vis(p_{ct})_h}{\partial \theta}], \end{aligned} \tag{20}$$

where θ is the set of CNN weights to be updated. Vis $(p_c)_h$ denotes the h-th element of $Vis(p_c)$.

3.4 Trip Planning

Trip Planning can be modeled using a bi-criteria generalization of travelling salesman problem (TSP) with two conflicting objectives: maximizing the collected utility and minimizing the travel cost. The orienteering problem (OP) is a variant of TSP that seeks for a trip that maximizes the total collected utility while maintaining the travel cost under a given value. That is, the travel cost objective is turned to a constraint. OP can be formulated as an integer programming problem as follows [15], [30]. Let m be the number of POIs, where the starting POI is denoted as l_1 and the destination POI is denoted as l_m. The utility of visiting POI l_i is represented by the popularity $Pop(l_i)$ and the user interest $Int(l_i)$ of this POI. The cost of traveling from l_i to l_j is calculated as the summation of the travelling time and the personalized visit duration of POI l_j. One main difference between our work and prior works is that we personalize the visit duration at each POI predicted by VPMF, instead of using the average visit duration for all users. With the time budget B, we want to find an itinerary $I = (l_1, \ldots, l_m)$ that satisfies the following constraints.

$$Max \sum_{i=2}^{M-1} \sum_{j=2}^{M} x_{i,j}(\eta Int(Cat_i) + (1 - \eta)Pop(i)) \tag{21}$$

$$\sum_{j=2}^{M} x_{1,j} = \sum_{i=1}^{M-1} x_{i,m} = 1 \tag{22}$$

$$\sum_{i=1}^{M-1} x_{i,k} = \sum_{j=2}^{M} x_{k,j} \leq 1, for\ all\ k = 2, \ldots, M-1 \tag{23}$$

$$\sum_{i=1}^{M-1} \sum_{j=2}^{M} Cost(i,j) x_{i,j} \leq B \tag{24}$$

$$2 \leq l_i \leq M, for\ all\ i = 2, \ldots, M \tag{25}$$

$$l_i - l_j + 1 \leq (M-1)(1 - x_{i,j}), for\ all\ i,j = 2, \ldots, M. \tag{26}$$

The objective function (i.e., Equation (21)) is to maximize the total popularity and the interest score of visited POIs in the trip, where η is the weight given to balance the popularity and the interest. For a path from l_1 to l_m, if POI l_i is followed by POI l_j, we set the variable $x_{i,j} = 1$. Otherwise, we set $x_{i,j} = 0$. Constraint (22) ensures that the trip starting at POI l_1 and ending at POI l_m. Constraint (23) ensures that the trip is connected and each POI is visited at most once. Constraint (24) ensures that the trip meets the time budget B, based on the function $Cost(l_i, l_j)$ that considers both the traveling time and the personalized POI visit duration. Constraints (25) and (26) ensure that there are no sub-tours in the proposed trip, adapted from the sub-tour elimination used in the travelling salesman problem [12]. The orienteering problem is NP-hard. Hence, exact solutions for the orienteering problem are not feasible for a large number of POIs. The orienteering problem can be formulated as an integer programming problem. For solving this integer programming problem, we use the lpsolve linear programming package [3] to obtain optimal solutions.

4 Experiments

4.1 Dataset

We apply the proposed photo2trip method on the Yahoo! Flickr Creative Commons 100M (YFCC100M) dataset [37],

TABLE 2
Dataset Description

City	# Images	# Users	# POI Visits	# Travel Sequences
Toronto	157,505	1,395	39,419	6,057
Budapest	145,364	954	18,513	2,361
Edinburgh	82,060	1,454	33,944	5,028
Vienna	461,905	1,155	34,515	3,193

the largest public multimedia collection released, which consists of 100 million photos and 0.8 million videos posted on Flickr with relevant meta information, such as the date/time taken, geo-location coordinates and geo-graphic accuracy. The geo-graphic accuracy ranges from the world level to the street level.

From this dataset, we use geo-tagged photos that were taken in four cities, namely Toronto, Budapest, Edinburgh, and Vienna. More details regarding this dataset are shown in Table 2. The dataset was previously used for tour recommendation by Lim et al. [25]. As described in [23], we first obtain a list of POIs from Wikipedia and then map these photos to user-POI visits. After that, we construct user travel sequences and evaluate our proposed approach.

4.2 Comparison Methods

In our experiments, we compare our proposed approaches with three popular baseline approaches and some recently proposed approach PersTour [25] and its extension [26], and POIRank [8]. A brief introduction of each of them is shown as follows.

- *Random Selection (Rand).* Iteratively and randomly choose a POI l_j from unvisited POIs as next POI.
- *Greedy Nearest (GNear).* Iteratively and greedily choose the nearest POI l_j with the least value $T^{Travel}(l_i, l_j)$ from unvisited POIs as next POI .
- *Greedy Most Popular (GPop).* Iteratively and greedily choose the most popular POI l_j with the most value $Pop(l_j)$ from unvisited POIs as next POI.
- *PersTour and $\eta = 0.5$ (PersTour-.5).* PersTour [25] with balanced emphasis on both POI popularity and user interest. That means the objective function is to maximize the total popularity and the interest score of POIs in the trip.
- *PersTour and $\eta = 1$ (PersTour-1).* PersTour [25] with full emphasis on user interest. In other words, the objective function is to maximize the total interest score of POIs in the trip.
- *PersTour using adaptive weighting by scaling (PersTour-AS).* PersTour [26] with emphasis on optimizing both POI popularity and time-based user interest with weighted updates. That means that the emphasis is based on adaptive weighting by scaling of POI visit counts.
- *PersTour using adaptive weighting by cumulative distribution (PersTour-AC).* PersTour [26] with full emphasis on optimizing time-based user interest with weighted updates, where emphasis is based on adaptive weighting by cumulative distribution of POI visit counts.
- *Photo2Trip and $\eta = 0.5$ (PT-VPMF-.5).* Photo2Trip based on VPMF [49] is proposed to integrate the visual features into the probabilistic matrix factorization model for better predicting user interests. The objective function is similar to PT-.5.
- *Photo2Trip and $\eta = 1$ (PT-VPMF-1).* Photo2Trip based on VPMF [49] with full emphasis user interest. And the objective function is the same as PT-1.

As described in Section 3.4, instead of using the average POI visit duration as user interest in PersTour [25], we choose the PMF [35] model to predict user visit interests in terms of different granularity. We first use the PMF model to predict the user visit interests on the category of a POI, and then we predict user visit interests on a specific POI. Our approaches are listed as follows.

- *Photo2Trip using PMF on POI Category level and $\eta = 0.5$ (PT-PMFC-.5).* Based on PersTour [25] with balanced emphasis on both POI popularity and user interests, we add the PMF model to predict user interests on a POI category. That means the prediction interests in one unvisited category is the same.
- *Photo2Trip using PMF on POI Category level and $\eta = 1$ (PT-PMFC-1).* Based on PersTour [25] with full emphasis on user interests, we add the PMF model to predict user interests on the category of a POI.
- *Photo2Trip using PMF on POI level and $\eta = 0.5$ (PT-PMF-.5).* Based on PersTour [25] with balanced emphasis on both POI popularity and user interests, we add the PMF model to predict user interest on a specific POI, more detail than the category level.
- *Photo2Trip using PMF on POI level and $\eta = 1$ (PT-PMF-1).* Based on PersTour [25] with full emphasis on user interests, we add the PMF model to predict user interest on a specific POI.

Again, the user-generated geo-tagged photos provide important contexts for predicting user visit interest for personalized tour recommendation. To integrate these photos into personalized tour recommendation, we incorporate the visual contents extracted only by CNN into the E2E-VPMF model. Since we noticed the advantage of predicting user visit interest on a specific POI, we use E2E-VPMF to predict user visit interest on a specific POI, respectively, instead of predicting user visit interest on the category of POIs. Therefore, we have following two approaches based on E2E-VPMF.

- *Photo2Trip using End-to-End VPMF on POI level and $\eta = 0.5$ (PT-E2E-VPMF-.5).* Based on VPMF [49] as described in Section 3.3, personalized tour recommendation uses the End-to-End VPMF model to predict user interest on a specific POI. The visual features of photos direct the learning process of latent user and POI factors. The objective function is to maximize the total popularity and the interest score of POIs in the trip.
- *Photo2Trip using End-to-End VPMF on POI level and $\eta = 1$ (PT-E2E-VPMF-1).* Based on VPMF [49], the objective function, in this case, is to maximize the total interest score of POIs in the personalized tour recommendation using the E2E-VPMF model to predict user interest on each POI.

TABLE 3
Performance Comparison of Tour Recommendation in Terms of Precision, Recall, and F1-Score on Four Datasets

Algo.	Toronto			Budapest			Edinburgh			Vienna		
	Pre.	*Rec.*	*F1-score*	*Pre.*	*Rec.*	*F1-score*	*Pre.*	*Rec.*	*F1-score*	*Pre.*	*Rec.*	*F1-score*
GNear	.464 ± .010	.544 ± .008	.484 ± .012	.359 ± .021	.477 ± .008	.393 ± .011	.386 ± .005	.501 ± .021	.422 ± .012	.385 ± .024	.530 ± .026	.426 ± .011
GPop	.611 ± .015	.389 ± .037	.466 ± .016	.544 ± .037	.350 ± .035	.413 ± .037	.592 ± .015	.459 ± .008	.503 ± .009	.543 ± .005	.364 ± .021	.423 ± .023
Rand	.451 ± .002	.274 ± .028	.336 ± .019	.401 ± .035	.237 ± .038	.289 ± .024	.450 ± .031	.271 ± .014	.325 ± .016	.487 ± .006	.285 ± .018	.351 ± .008
PersTour-1	.720 ± .015	.755 ± .021	.728 ± .018	.772 ± .021	.777 ± .018	.768 ± .031	.604 ± .020	.662 ± .011	.616 ± .029	.618 ± .002	.660 ± .013	.625 ± .014
PersTour-.5	.704 ± .014	.774 ± .025	.729 ± .011	.781 ± .009	.788 ± .010	.777 ± .008	.631 ± .014	.742 ± .019	.671 ± .014	.646 ± .006	.715 ± .009	.666 ± .006
PersTour-AS	.710 ± .012	.767 ± .011	.722 ± .014	.775 ± .025	.787 ± .020	.769 ± .027	.625 ± .015	.736 ± .007	.629 ± .019	.631 ± .010	.714 ± .008	.652 ± .012
PersTour-AC	.708 ± .013	.766 ± .019	.734 ± .011	.787 ± .011	.798 ± .008	.783 ± .009	.623 ± .013	.723 ± .011	.630 ± .012	.652 ± .011	.710 ± .009	.665 ± .010
PT-PMFC-1	.724 ± .021	.755 ± .024	.731 ± .019	.807 ± .012	.792 ± .018	.792 ± .020	.605 ± .017	.663 ± .018	.618 ± .021	.631 ± .019	.664 ± .020	.635 ± .021
PT-PMFC-.5	.718 ± .011	.779 ± .015	.739 ± .020	.816 ± .020	.801 ± .035	.801 ± .031	.640 ± .007	.750 ± .009	.680 ± .010	.645 ± .012	.715 ± .020	.667 ± .025
PT-PMF-1	.746 ± .011	.769 ± .012	.751 ± .009	.813 ± .021	.797 ± .026	.795 ± .031	.620 ± .025	.674 ± .016	.631 ± .035	.654 ± .008	.676 ± .012	.651 ± .015
PT-PMF-.5	.725 ± .012	.791 ± .015	.749 ± .021	.821 ± .025	.806 ± .021	.803 ± .030	.643 ± .009	.756 ± .012	.685 ± .011	.655 ± .017	.725 ± .015	.676 ± .020
PT-VPMF-1	.749 ± .021	.805 ± .011	.765 ± .019	.812 ± .002	.808 ± .012	.809 ± .007	.621 ± .015	.678 ± .013	.634 ± .021	.660 ± .008	.685 ± .028	.676 ± .019
PT-VPMF-.5	.728 ± .022	.828 ± .023	.762 ± .029	.831 ± .011	.809 ± .012	.819 ± .023	.645 ± .025	.768 ± .016	.696 ± .035	.672 ± .011	.751 ± .009	.709 ± .023
PT-E2E-VPMF-1	**.753** ± **.015**	**.813** ± **.008**	**.780** ± **.013**	**.824** ± **.016**	**.817** ± **.011**	**.820** ± **.009**	**.631** ± **.013**	**.702** ± **.007**	**.665** ± **.018**	**.670** ± **.006**	**.697** ± **.023**	**.683** ± **.015**
PT-E2E-VPMF-.5	**.731** ± **.018**	**.835** ± **.015**	**.779** ± **.023**	**.835** ± **.007**	**.812** ± **.012**	**.823** ± **.019**	**.650** ± **.021**	**.783** ± **.011**	**.710** ± **.025**	**.679** ± **.008**	**.764** ± **.011**	**.719** ± **.009**
Improv (vs. PersTour-1)	4.58%	7.68%	7.14%	6.74%	5.15%	6.77%	4.47%	6.04%	7.95%	8.41%	5.61%	9.28%
Improv (vs. PersTour-.5)	3.84%	7.88%	6.86%	6.91%	3.05%	5.92%	3.01%	5.53%	5.81%	5.11%	6.85%	7.96%

The best performance in each column is boldfaced (higher is better). Improvements over PersTour-1 and PersTour-.5 are shown in the last two rows.

4.3 Evaluation Metrics

We evaluate the popular baseline approaches and our proposed E2E-VPMF framework, using leave-one-out cross-validation [20]. When evaluating a specific travel sequence of a user, we use the user's other travel sequences as training data. At last, we evaluate the performance of each algorithm using the following metrics.

- *Tour Precision*. The precision of POIs recommended in the trip is the proportion of POIs recommended in a trip that was also in a user's real-life travel sequence, defined as $\frac{|P_r \cap P_v|}{|P_r|}$, where P_r and P_v are the set of POIs recommended in the tour and visited by the user in real-life, respectively.
- *Tour Recall*. The recall of POI recommendation in the trip is the proportion of POIs in a user's real-life travel sequence that was recommended, defined as $\frac{|P_r \cap P_v|}{|P_v|}$, where P_r and P_v are the set of POIs recommended in the trip and visited by the user in his/her real-life travel sequence, respectively.
- *Trip F_1-score*. It combines both precision and recall of a recommended trip together with the harmonic mean.
- *Root-Mean-Square Error (RMSE) of POI Visit Duration*. RMSE is a frequently used to measure the difference between a value predicted by a model and the value actually observed. Let p be a POI in recommended itinerary I, which was visited in real-life. Let D_r be the recommended duration and D_v be the duration in real-life respectively. Then, RMSE is defined as follows.

$$RMSE = \sqrt{\frac{\sum_{p \in I}(D_r - D_v)^2}{|I|}}.$$

4.4 Parameter Settings

For each method, there are some hyper-parameters to tune. We consider the regularization term from $\{0.0001, 0.001, 0.01, 0.1, 1\}$, the latent dimension D ranging from $\{10, 20, 30, 40, 50\}$, and the learning rate from $\{0.001, 0.005, 0.01, 0.1, 1\}$. In our proposed framework (i.e., E2E-VPMF), we set the dimension of the latent space D as 50, considering both efficiency and effectiveness. The regularization parameters λ_1 and λ_2 of E2E-VPMF are chosen by cross-validation, and finally set at 0.01. For the image weighting parameter α and the number of negative examples L in the negative sampling process, we will discuss their impact of different settings in Section 4.6. As a preprocessing step, we use a pretrained VGG16 model on ImageNet to initialize the weights. In addition, we report the result of each method under its optimal setting of their hyper-parameters.

4.5 Results and Discussion

In this part, we discuss the experimental results of our proposed model with all baselines on four datasets. First we analysis the overall performance, and then we construct detailed analysis of the results of different models.

4.5.1 Overall Performance

Table 3 presents an overview of results in terms of Precision, Recall and F1-score across all four datasets. The results show that all variants of our Photo2Trip framework outperform the three greedy-based baselines (i.e., GNear, GPop and Rand). This indicates the importance of visual contents of geo-tagged photos. PersTour (i.e., PersTour-1 and PersTour-.5) and its extended version (i.e., PersTour-AS and PersTour-AC) achieved a comparable performance. This observation illustrates the effectiveness of time-based user interest in recommending travel routes. Finally, for our model, we can see that both PT-E2E-VPMF-1 and PT-E2E-VPMF-.5 consistently outperform the baseline algorithms on all datasets in terms of Precision, Recall and F1-score. This suggests that the end-to-end training framework with incorporating visual content of images into probabilistic matrix factorization can significantly improve the performance of tour recommendation.

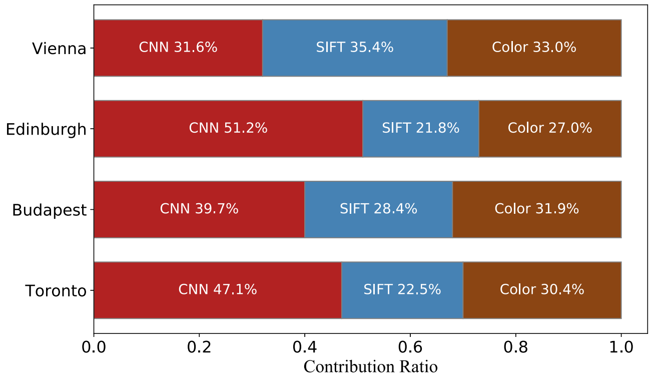


Fig. 5. Performance improvement contribution of different visual features in E2E-VPMF.

4.5.2 Effectiveness of PMF on the Category Level of POIs.

We first evaluate the performance of incorporating PMF into trip planning to predict user visit interests on the category level of POIs. As shown in Table 3, both PT-PMFC-1 and PT-PMFC-.5 in most cases outperform the state-of-the-art PersTour, in terms of precision, recall and F_1. As expected, PT-PMFC consistently outperforms the greedy and random methods. This observation shows the effectiveness of integrating collaborative filtering into predicting visit interests in trip planning, which indicates that using the PMF model to predict user interests on the category level is more accurate than using the average visit time of all user in a category as user interest.

4.5.3 Effectiveness of PMF on the POI Level

We then evaluate the performance of incorporating PMF into trip planning to predict user visit interests on the POI level, a lower granular level than the category level. As shown in Table 3, both PT-PMF-1 and PT-PMF-.5 consistently outperform both PT-PMFC-1 and PT-PMFC-.5, in terms of precision, recall and F_1. The results indicate that predicting user visit interests on the POI level is more accurate and more effectiveness in trip planning, comparing with predicting on the category level of POIs.

4.5.4 Effectiveness of VPMF on the POI Level

We further evaluate the performance of integrating VPMF into trip planning to predict user visit interests on the POI level by leverage visual content in geo-tagged photos. As shown in Table 3, both PT-VPMF-1 and PT-VPMF-.5 consistently outperform both PT-PMF-1 and PT-PMF-.5, in terms of precision, recall and F_1. The results indicate that predicting user interests by integrating visual content inside the PMF model is more accurate than the approaches based on the PMF model, and show significant effectiveness in trip planning. Overall, PT-VPMF-1 outperforms the existing popular approach PT-1 5.38 percent, and PT-VPMF-.5 outperforms the existing popular approach PT-.5 5.03 percent with respect to the average F_1 value on the two cities.

4.5.5 Effectiveness of E2E-VPMF on the POI Level

We evaluate the performance of integrating E2E-VPMF into trip planning to predict user visit interests on the POI level by modeling and extracting high-level visual contents through CNN. The E2E-VPMF model uses user-generated photos to help study the latent factors of users and POIs, which indirectly influence the preference scores and is therefore more robust to noises. Table 3 shows that PT-E2E-VPMF-1 and PT-E2E-VPMF-.5 consistently outperform PT-PMF-1 and PT-PMF-.5, respectively, in terms of precision, recall and F_1. Specifically, PT-E2E-VPMF-1 outperforms the existing popular approach PT-1 7.79 percent on average, and PT-E2E-VPMF-.5 outperforms the existing popular approach PT-.5 6.64 percent on average with respect to the average F_1 value on the two cities.

Moreover, we can find that PT-E2E-VPMF performs better than PT-VPMF in terms of Precision, Recall and F1-score on all datasets. Because images uploaded by users and associated with POIs contain rich information about user preferences and POI properties, different from VPMF, E2E-VPMF first incorporates visual contents of photos into a probabilistic matrix factorization model, and then form an end-to-end training framework for guiding the learning process of user and POI latent factors. These results further suggest that the effectiveness of incorporating visual contents for POI recommendation, which can improve the tour recommendation performance.

4.5.6 Effectiveness of Different Visual Contents on E2E-VPMF

Similarly, we also have added two traditional visual features to the E2E-VPMF model for tour recommendation, i.e., SIFT and Color Histogram, as described in Section 2.5. The contributions of different visual features in our proposed E2E-VPMF model are shown in Fig. 5. We can observe that CNN visual contents are the best on three datasets. This may be because we use the pre-trained CNN model on ImageNet, which is able to capture high-order visual features of photos, yet SIFT and Color are manually crafted low-level features, which are not so discriminative. In addition, SIFT is better than other two features on the Vienna dataset. This demonstrates that combination of high and low-order features can achieve better results.

4.5.7 Visiting Duration Prediction Accuracy

With the availability of user interest predictions, we can personalize the POI visit duration more accuracy for each user. Apart from the accuracy of POIs recommended in a trip, recommending the appropriate amount of time to spend at a specific POI is another important consideration in tour recommendation. Visit duration at each POI is important in trip planning. In general, users intend to spend less time on uninteresting POIs to save time budget for interesting POIs. This matches user's behaviors that users usually prefer visiting a few POIs with high interest using all time budget to visiting many POIs with less interest. As shown in Table 4, the recommended personalized POI visit duration of PT-VPMF outperforms state-of-the-art personalized methods PT over 10 percent in all case and over

TABLE 4
Performance Comparison of Visiting Duration Prediction in Terms of RMSE

Algo.	*RMSE*			
	Toronto	Budapest	Edinburgh	Vienna
PersTour-1	145.20 ± 9.25	65.35 ± 6.31	73.39 ± 9.53	62.99 ± 5.28
PersTour-.5	143.55 ± 9.88	57.27 ± 5.12	91.48 ± 5.07	68.93 ± 5.69
PT-PMFC-1	127.29 ± 7.14	52.53 ± 5.01	70.17 ± 4.52	59.90 ± 6.04
PT-PMFC-.5	121.87 ± 8.59	50.52 ± 8.25	84.23 ± 9.35	61.26 ± 6.28
PT-PMF-1	110.90 ± 9.99	44.19 ± 9.18	66.68 ± 5.35	52.47 ± 5.87
PT-PMF-.5	104.67 ± 6.78	47.37 ± 9.21	**73.72 ± 8.53**	51.31 ± 6.21
PT-VPMF-1	109.76 ± 6.51	32.71 ± 5.35	65.72 ± 8.05	48.61 ± 7.25
PT-VPMF-.5	101.85 ± 7.68	41.87 ± 8.38	82.12 ± 9.94	44.88 ± 6.01
PT-E2E-VPMF-1	**102.57 ± 5.89**	**23.36 ± 5.03**	**62.28 ± 7.65**	**45.80 ± 6.73**
PT-E2E-VPMF-.5	**97.43 ± 7.38**	**34.05 ± 7.98**	77.52 ± 9.03	**43.65 ± 5.85**
PT-VPMF-1 over PT-1	24.41%	49.95%	10.45%	22.83%
PT-VPMF-.5 over PT-.5	29.04%	26.89%	17.76%	34.89%
PT-E2E-VPMF-1 over PT-1	29.36%	64.25%	15.14%	27.29%
PT-E2E-VPMF-.5 over PT-.5	32.13%	40.54%	18.01%	36.67%

TABLE 5
Performance Comparison with Cold Start Scenario in Terms of Precision, Recall, and F1-Score

Algo.	Toronto			Budapest			Edinburgh			Vienna		
	Pre.	*Rec.*	*F1-score*	*Pre.*	*Rec.*	*F1-score*	*Pre.*	*Rec.*	*F1-score*	*Pre.*	*Rec.*	*F1-score*
PersTour-1	.678 ± .004	.682 ± .011	.672 ± .002	.572 ± .020	.582 ± .008	.567 ± .017	.522 ± .010	.584 ± .006	.539 ± .019	.602 ± .012	.566 ± .011	.567 ± .012
PersTour-.5	.635 ± .004	.741 ± .015	.676 ± .011	.488 ± .004	.681 ± .011	.548 ± .018	.522 ± .011	.703 ± .032	.588 ± .014	.486 ± .003	.630 ± .009	.533 ± .023
PT-VPMF-1	.703 ± .012	.703 ± .011	.695 ± .009	.611 ± .008	.607 ± .011	.596 ± .012	.580 ± .004	.607 ± .018	.593 ± .012	.653 ± .009	.584 ± .014	.592 ± .011
PT-VPMF-.5	.691 ± .023	.808 ± .015	.736 ± .027	.528 ± .014	.677 ± .010	.573 ± .023	.583 ± .018	.722 ± .002	.630 ± .005	.484 ± .017	.662 ± .024	.524 ± .012
PT-E2E-VPMF-1	**.714 ± .008**	**.709 ± .011**	**.711 ± .007**	**.625 ± .012**	**.611 ± .008**	**.618 ± .015**	**.597 ± .006**	**.609 ± .015**	**.603 ± .009**	**.671 ± .011**	**.598 ± .017**	**.632 ± .013**
PT-E2E-VPMF-.5	**.699 ± .018**	**.814 ± .014**	**.752 ± .021**	**.536 ± .009**	**.690** ± .008	**.603 ± .025**	**.589 ± .013**	**.724 ± .005**	**.650 ± .007**	**.499 ± .017**	**.673 ± .019**	**.553 ± .014**

27 percent on average in terms of RMSE. PT-E2E-VPMF outperforms state-of-the-art personalized methods PT over 15 percent in all case and over 32 percent on average in terms of RMSE. This shows that personalized user visit duration prediction at a POI using VPMF more accurately reflects the real-life POI visit duration of users.

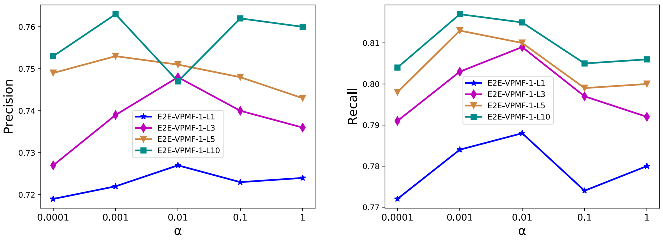
4.5.8 Cold Start Scenario

A cold start user means a user without any travel history data. To investigate the performance of VPMF and E2E-VPMF for cold start users, we adapted the concept of leave-one-out cross-validation [20] in our experiments. That is, we leave one user out for testing. Specifically, we removed all historical travel data of this user and only kept his/her photos with all geo-tags removed. As we lack the check-in history of this user, this user is considered as a cold start user. Only visual content in photos can help reveal user interest. Therefore, the model must have the ability to address the inherent cold start nature and to recommend trip plan accurately to achieve acceptable performance. As shown in Table 5, the performance of all methods decreases comparing with warm start shown in Table 3. PT-PMF has no results in Table 5 since it cannot handle cold start users. On this cold start scenario, the performance reduction of the proposed framework VPMF and E2E-VPMF are much smaller compared to the PT, PT-VPMF-1 outperforms PT-1 4.41 percent, and PT-VPMF-.5 outperforms PT-.5 6.73 percent with respect to the average F_1 value. Moreover, PT-E2E-VPMF-1 outperforms PT-1 7.40 percent, and PT-E2E-VPMF-.5 outperforms PT-.5 9.13 percent with respect to the average F_1 value. The results support that the proposed framework by incorporating visual features of photos can alleviate the cold start problem for tour recommendation.

4.6 Influence of Hyper-Parameters

In this section, we explore the influence of two important parameters on the performance of E2E-VPMF: α and L, where α dominates the importance of photos in studying the latent factors of users and POIs, and L controls the accuracy of Equations (12) and (13) in approximating Equations (7) and (8) respectively. We apply a grid search over the combination of varying $\alpha \in \{0.0001, 0.001, 0.01, 0.1, 1\}$ and $L \in \{1, 3, 5, 10\}$ for accurate recommendation. The results are shown in Fig. 6. Due to limitation of space, we only show the changes of Precision and Recall of E2E-VPMF-1 on the Toronto dataset. And we have obtained similar results on other datasets in terms of three metrics.

On one hand, we can observe that as L grows from 1 to 10, the performance of E2E-VPMF increases gradually on most cases, which is in line with result in [33]. It indicates that a larger L can achieve better performance and also



(a) Precision - Toronto (b) Recall - Toronto

Fig. 6. The performance of E2E-VPMF-1 with varying L and α in terms of Precision and recall on Toronto dataset.

means more computing costs. Therefore, there is a balance between efficiency and effectiveness. On the other hand, we notice that the proposed framework E2E-VPMF comes down to PMF when setting α to 0. When increasing α , we incorporate visual features into PMF for tour recommendation. And with the increase of α, the performance of E2E-VPMF is improved first and then deteriorated on two cases. When setting $\alpha = 0.001$ and $L = 10$, the proposed framework E2E-VPMF achieves the best performance in terms of precision and recall. In general, these results again highlight the importance of photos for tour recommendation.

5 Related Work

This paper makes a forward step for tour recommendation, which is rooted in POI recommendation. POI recommendation is to recommend a list of top k most relevant POIs to a user, based on user implicit feedback, such as check-in frequency. Collaborative filtering is widely used in POI recommendation. The state-of-the-art collaborative filtering (CF) is based on matrix factorization and its variants [21], [22], [35], [43]. Salakhutdinov & Mnih [35] proposed a PMF model in a Bayesian probabilistic framework to include Gaussian noise in observations. Under the Gaussian assumption, maximizing the posterior probability over latent features is equivalent to minimizing the square error.

Recently, more advanced models have been proposed to exploit additional information for POI recommendation [1], [17], [44], [48], such as check-in locations, social influence, temporal information and transition between POIs.

Ye et al. [45], [46] considered the social influence under the framework of a user-based CF model and modeled the geographical influence by a Bayesian CF model. Moreover, both Yuan et al. [47] and Gao et al. [13] introduced temporal preference to enhance the efficiency and effectiveness of Ye et al.'s solution. Cheng et al. [10] considered more comprehensive information, such as the multi-center of user check-in patterns, and skewed user check-in frequency. Moreover, Liu et al. [29] proposed a bi-weighted low-rank graph construction model, which integrates users' interests and their evolving sequential preferences with temporal interval assessment to provide POI recommendations for a specific time period. Jiang et al. [19] proposed an author topic model-based collaborative filtering method for POI recommendation, which employs user preference topics, such as cultural, cityscape, or landmark, extracted from the textual description of photos via the author topic model. However, most of these methods did not explicitly consider the visual content in user-generated photos in POI recommendation. Besides, they evaluated each venue independently without considering other information and ignored the order of visits. Moreover, there are no overall time constraints, and traveling time is not considered. In this paper, we focus on tour recommendation which recommends relevant POIs as well as order the POIs into a trip to satisfy different constraints, e.g., the maximum travel time budget.

Tour recommendation has become very important in recent years. A large number of public available traveler e-footprints (such as geo-tagged photos and blogs) make automatic trip planning possible. Arase et al. [2] proposed a photo trip pattern mining framework to detect users' frequent trip patterns extracted from public geo-tagged photos, i.e., typical sequences of visited cities and visit duration as well as trip themes that characterize the trip patterns. Lu et al. [32] leveraged existing travel clues from geo-tagged photos to suggest customized route plans according to users' preferences. They used geo-tagged photos to discover the tour paths within a destination and travel routes between destinations. Cheng et al. [9] further proposed a probability-based personalize travel recommendation model based on user's profiles (such as gender, age, and race) by leveraging users' attributes in user-generated photos. Lim et al. [24] utilized geo-tagged photos to derive attraction popularity, user interests and queuing times, which recommend personalized and queue-aware itineraries. Bin et al. [4] integrated multi-source tourism big data on websites to generate POIs knowledgebase and POIs visit sequences, and then designed the POIs travel route recommendation method under tourist personal constraints. He et al. [18] proposed a POI embedding model to jointly learn the impact of recommendation contextual factors, including the POI popularity, other POIs co-occurring in the trip and the preferences of the user, for trip recommendation. Moreover, Cai et al. [7] proposed an itinerary recommender system with semantic trajectory pattern mining from geo-tagged photos, considering spatio-temporal, spatial semantics dimensions, and so on, to customize user requests. To further satisfy tourists' demand, Gaonkar et al. [14] leveraged social media, more explicitly photo uploads and their tags, to reverse engineer historical user itineraries, which converges to an individual trip that is tailored to an user's interest. Although they also utilized visual features in geo-tagged photos, they only used facial visual content to infer user's profiles, and did not take advantage of general visual features.

A recent work named PersTour [25], [27] is closely related to our work and reflects the levels of user's interest based on visit durations, which are obtained from real-life travel sequences based on geo-tagged photos. PersTour uses average POI visit duration as user interest and has not employed collaborative filtering to predict user interest. However, the major difference between our work and related research described above is that we extract visual features from user-generated photos and consider these visual features with the mobility pattern of tourists to help learn the latent features of both users and POIs for the task of our personalized tour recommendation.

6 Conclusion

In this paper, a tour recommender system leveraging geo-tagged photos, named 'Photo2Trip', was proposed to recommend not only suitable POIs to visit but also visit duration at each POI. Specifically, we proposed a Visual-enhanced Probabilistic Matrix Factorization model, which integrated visual features into the collaborative filtering model, to learn user interests by leveraging the historical travel records. Our work improved existing tour recommendation research in three ways: (i) we introduced collaborative filtering into trip planning to predict user visit preferences of non-visited POIs, instead of using the average visit duration of each category of POIs for all users as individual interest; (ii) we extracted and integrated visual features in user-generated photos of POIs into the collaborative filtering model PMF to further improve user interest prediction; (iii) we further proposed an E2E-VPMF training framework to improve the performance of POI recommendation significantly, which makes visual features of images participate in the learning of user (item) latent vectors.

Using the Yahoo! Flickr dataset across four cities, we evaluated the effectiveness of our proposed approach against various baseline methods. The extensive experimental results showed that: (i) using collaborative filtering to predict user interest resulted in accurate prediction to the real-life travel sequences of users, in terms of both precision and F_1-score; (ii) incorporating visual features into the PMF model could further improve the accuracy of prediction; (iii) our proposed VPMF approaches predicted personalized POI visit duration more accurately; and (iv) incorporating visual features into PMF significantly alleviated the cold start problem.

Acknowledgments

This research was supported by NSFC (No. 61876117, 61876217, 61772356, 6172820561972069).

References

[1] R. P. Adams, G. E. Dahl, and I. Murray, “Incorporating side information in probabilistic matrix factorization with gaussian processes,” in *Proc. 26th Conf. Uncertainty Artificial Intell.*, 2010, pp. 1–9.

[2] Y. Arase, X. Xie, T. Hara, and S. Nishio, “Mining people’s trips from large scale geo-tagged photos,” in *Proc. 18th ACM Int. Conf. Multimedia.*, 2010, pp. 133–142.

[3] M. Berkelaar, K. Eikland, P. Notebaert, et al., “lpsolve: Open source (mixed-integer) linear programming system,” *Eindhoven U. Technol.*, 2004.

[4] C. Bin, T. Gu, Y. Sun, L. Chang, W. Sun, and L. Sun, “Personalized pois travel route recommendation system based on tourism big data,” in *Proc. Pacific Rim Int. Conf. Artificial Intell.*, 2018, pp. 290–299.

[5] I. Brilhante, J. A. Macedo, F. M. Nardini, R. Perego, and C. Renso, “Where shall we go today?: Planning touristic tours with tripbuilder,” in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manag.*, 2013, pp. 757–762.

[6] I. R. Brilhante, J. A. Macedo, F. M. Nardini, R. Perego, and C. Renso, “On planning sightseeing tours with tripbuilder,” *Inf. Process. Manage.*, vol. 51, no. 2, pp. 1–15, 2015.

[7] G. Cai, K. Lee, and I. Lee, “Itinerary recommender system with semantic trajectory pattern mining from geo-tagged photos,” *Expert Syst. Appl.*, vol. 94, pp. 32–40, 2018.

[8] D. Chen, C. S. Ong, and L. Xie, “Learning points and routes to recommend trajectories,” in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 2227–2232.

[9] A.-J. Cheng, Y.-Y. Chen, Y.-T. Huang, W. H. Hsu, and H.-Y. M. Liao, “Personalized travel recommendation by mining people attributes from community-contributed photos,” in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 83–92.

[10] C. Cheng, H. Yang, I. King, and M. R. Lyu, “Fused matrix factorization with geographical and social influence in location-based social networks.” in *Proc. 26th AAAI Conf. Artificial Intell.*, 2012, vol. 12, Art. no. 1.

[11] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual categorization with bags of keypoints,” in *Proc. Eur. Conf. Comput. Vis.*, 2004, vol. 1, pp. 1–22.

[12] D. Feillet, P. Dejax, and M. Gendreau, “Traveling salesman problems with profits,” *Transp. Sci.*, vol. 39, no. 2, pp. 188–205, 2005.

[13] H. Gao, J. Tang, X. Hu, and H. Liu, “Exploring temporal effects for location recommendation on location-based social networks,” in *Proc. 7th ACM Conf. Recommender Syst.*, 2013, pp. 93–100.

[14] R. Gaonkar, M. Tavakol, and U. Brefeld, “Mdp-based itinerary recommendation using geo-tagged social media,” in *Proc. Int. Symp. Intell. Data Anal.*, 2018, pp. 111–123.

[15] D. Gavalas, C. Konstantopoulos, K. Mastakas, and G. Pantziou, “A survey on algorithmic approaches for solving tourist trip design problems,” *J. Heuristics*, vol. 20, no. 3, pp. 291–328, 2014.

[16] Y. Ge, Q. Liu, H. Xiong, A. Tuzhilin, and J. Chen, “Cost-aware travel tour recommendation,” in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 983–991.

[17] Q. Gu, J. Zhou, and C. Ding, “Collaborative filtering: Weighted nonnegative matrix factorization incorporating user and item graphs,” in *Proc. SIAM Int. Conf. Data Mining*, 2010, pp. 199–210.

[18] J. He, J. Qi, and K. Ramamohanarao, “A jointly learned context-aware place of interest embedding for trip recommendations,” *CoRR*, vol. abs/1808.08023, 2018.

[19] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei, “Author topic model-based collaborative filtering for personalized poi recommendations,” *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 907–918, Jun. 2015.

[20] R. Kohavi, et al.,“A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proc. 14th Int. Joint Conf. Artificial Intell.* - Vol. 2, 1995, vol. 14, pp. 1137–1145.

[21] Y. Koren, “Factorization meets the neighborhood: A multifaceted collaborative filtering model,” in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 426–434.

[22] Y. Koren, R. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” *Comput.*, vol. 42, no. 8, pp. 30–37, 2009.

[23] K. H. Lim, “Recommending tours and places-of-interest based on user interests from geo-tagged photos,” in *Proc. ACM SIGMOD PhD Symp.*, 2015, pp. 33–38.

[24] K. H. Lim, J. Chan, S. Karunasekera, and C. Leckie, “Personalized itinerary recommendation with queuing time awareness,” in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2017, pp. 325–334.

[25] K. H. Lim, J. Chan, C. Leckie, and S. Karunasekera, “Personalized tour recommendation based on user interests and points of interest visit durations,” in *Proc. 24th Int. Conf. Artificial Intell.*, 2015, pp. 1778–1784.

[26] K. H. Lim, J. Chan, C. Leckie, and S. Karunasekera, “Personalized trip recommendation for tourists based on user interests, points of interest visit durations and visit recency,” *Knowl. Inf. Syst.*, vol. 54, no. 2, pp. 375–406, 2018.

[27] K. H. Lim, J. Chan, C. Leckie, and S. Karunasekera, “Personalized trip recommendation for tourists based on user interests, points of interest visit durations and visit recency,” *Knowl. Inf. Syst.*, vol. 54, no. 2, pp. 375–406, 2018.

[28] B. Liu and H. Xiong, “Point-of-interest recommendation in location based social networks with topic and location awareness,” in *Proc. SIAM Int. Conf. Data Mining*, 2013, pp. 396–404.

[29] Y. Liu, C. Liu, B. Liu, M. Qu, and H. Xiong, “Unified point-of-interest recommendation with temporal interval assessment,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1015–1024.

[30] Y. Liu, C. Liu, N. J. Yuan, L. Duan, Y. Fu, H. Xiong, S. Xu, and J. Wu, “Exploiting heterogeneous human mobility patterns for intelligent bus routing,” in *Proc. IEEE Int. Conf. Data Mining*, 2014, pp. 360–369.

[31] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, 1999, vol. 2, pp. 1150–1157.

[32] X. Lu, C. Wang, J.-M. Yang, Y. Pang, and L. Zhang, “Photo2trip: Generating travel routes from geo-tagged photos for trip planning,” in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 143–152.

[33] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.* - Vol. 2, 2013, pp. 3111–3119.

[34] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “Cnn features off-the-shelf: an astounding baseline for recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2014, pp. 512–519.

[35] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proc. 20th Int. Conf. Neural Inf. Process. Syst.*, 2007, vol. 1, pp. 1257–1264.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.

[37] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "The new data and new challenges in multimedia research," *arXiv:1503.01817*, vol. 1, no. 8, 2015.

[38] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu, "Multimodal graph-based reranking for web image search," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4649–4661, Nov. 2012.

[39] M. Wang, C. Luo, B. Ni, J. Yuan, J. Wang, and S. Yan, "First-person daily activity recognition with manipulated object proposals and non-linear feature fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2946–2955, Oct. 2018.

[40] S. Wang, J. Tang, C. Aggarwal, and H. Liu, "Linked document embedding for classification," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 115–124.

[41] S. Wang, Y. Wang, J. Tang, K. Shu, S. Ranganath, and H. Liu, "What your images reveal: Exploiting visual contents for point-of-interest recommendation," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 391–400.

[42] X. Wang and Y. Wang, "Improving content-based and hybrid music recommendation using deep learning," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 627–636.

[43] L. Wu, E. Chen, Q. Liu, L. Xu, T. Bao, and L. Zhang, "Leveraging tagging for neighborhood-aware probabilistic matrix factorization," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 1854–1858.

[44] B. Xia, Y. Li, Q. Li, and T. Li, "Attention-based recurrent neural network for location recommendation," in *Proc. 12th Int. Conf. Intell. Syst. Knowl. Eng.*, 2017, pp. 1–6.

[45] M. Ye, P. Yin, and W.-C. Lee, "Location recommendation for location-based social networks," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2010, pp. 458–461.

[46] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee, "Exploiting geographical influence for collaborative point-of-interest recommendation," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2011, pp. 325–334.

[47] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann, "Time-aware point-of-interest recommendation," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2013, pp. 363–372.

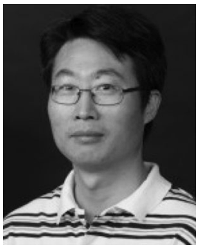
[48] Z. Zhang, Y. Liu, Z. Zhang, and B. Shen, "Fused matrix factorization with multi-tag, social and geographical influences for poi recommendation," *World Wide Web*, vol. 22, pp. 1135–1150, May 2018.

[49] P. Zhao, X. Xu, Y. Liu, V. S. Sheng, K. Zheng, and H. Xiong, "Photo2trip: Exploiting visual contents in geo-tagged photos for personalized tour recommendation," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 916–924.

Pengpeng Zhao received the PhD degree in computer science from Soochow University, in 2008. He is a professor with the School of Computer Science and Technology, Soochow University. From 2016 to 2017, he was a visiting scholar, working with the Data Mining and Business Analysis Laboratory, Rutgers University. His current research interests include data mining, deep learning, big data analysis, and recommender systems. He has published more than 60 papers in prestigious international conferences and journals, including ACM MM, AAAI, IJCAI, ICDM, CIKM, DASFAA, and ICME. He was a program committee member of international conferences, such as AAAI, IJCAI, CIKM, and PAKDD.

Chengfeng Xu is currently working toward the MS degree in the School of Computer Science and Technology, Soochow University, Suzhou. Her main research interests include spatial data processing, recommender systems, and data mining. She has published one paper in WWW2019 and one paper in IJCAI2019.

Yanchi Liu received the PhD degree in information technology from Rutgers, the State University of New Jersey, and the PhD degree in management science and engineering from the University of Science and Technology Beijing. He is a researcher with NEC Labs America. His research interests include data mining, business intelligence, urban computing, and recommender systems. He has published in the *IEEE Transactions on Cybernetics*, the *ACM Transactions on Intelligent Systems and Technology*, KDD, IJCAI, ICDM, etc.



Victor S. Sheng received the master's degree in computer science from the University of New Brunswick, Canada, in 2003, and the PhD degree in computer science from Western University, Ontario, Canada, in 2007. He is an associate professor of computer science with Texas Tech University, and the founding director of the Data Analytics Lab (DAL). His research interests include data mining, machine learning, crowdsourcing, and related applications in business, industry, medical informatics, and software engineering. He is a senior member of the IEEE and a lifetime member of the ACM, and a SPC and PC member for many international conferences.

Kai Zheng received the PhD degree in computer science from The University of Queensland, in 2012. He is a professor of computer science with the University of Electronic Science and Technology of China. He has been working in the area of spatial-temporal databases, uncertain databases, social-media analysis, inmemory computing, and blockchain technologies. He has published more than 100 papers in prestigious journals and conferences in data management field such as SIGMOD, ICDE, the *VLDB Journal*, the ACM Transactions, and IEEE Transactions. He is a member of the IEEE.

Hui Xiong received the BE degree from the University of Science and Technology of China (USTC), China, the MS degree from the National University of Singapore (NUS), Singapore, and the PhD degree from the University of Minnesota (UMN). He is currently a full professor and vice chair of the Management Science and Information Systems Department, and the director of the Rutgers Center for Information Assurance at Rutgers, the State University of New Jersey. His general area of research is data and knowledge engineering, with a focus on developing effective and efficient data analysis techniques for emerging data intensive applications. He is an ACM distinguished scientist and a senior member of the IEEE.

Xiaofang Zhou received the bachelor's and master's degrees in computer science from Nanjing University, in 1984 and 1987, respectively, and the PhD degree in computer science from the University of Queensland, in 1994. He is a professor of computer science with the University of Queensland. He is the head of the Data and Knowledge Engineering Research Division, School of Information Technology and Electrical Engineering. He is also a specially appointed adjunct professor with Soochow University, China. His research is focused on finding effective and efficient solutions to managing integrating, and analyzing very large amounts of complex data for business and scientific applications. His research interests include spatial and multimedia databases, high performance query processing, web information systems, data mining, and data quality management. He is a fellow of the IEEE.